



Leveraging Large Language Models for Effective Data Augmentation in Imbalanced Rumor and Fake-News Detection

Mojgan Askarizade^{a,*}

^a*Department of Computer Engineering, Faculty of Engineering, Ardakan University, Ardakan, Yazd, Iran.*

ARTICLE INFO.

Article history:

Received: 27 April 2026

Revised: 15 September 2026

Accepted: 19 October 2026

Published Online: 24 October 2026

Keywords:

Rumor Detection, Data Augmentation, Large Language Models (LLMs), Class Imbalance, Misinformation.

ABSTRACT

Rumor detection on social media remains challenging due to rapidly spreading information and the prevalence of misleading or unverified content. A major obstacle is the imbalance in available datasets, where minority classes, such as verified rumors, often represent less than 30% of samples. This imbalance biases classifiers toward majority classes and limits accurate misinformation detection. To address this, we propose a multi-model data augmentation framework leveraging four large language models; DeepSeek-R1-Distill-Qwen-1.5B, Gemma-2-2B, GPT-4o-mini, and LLaMA-3.2-3B fine-tuned on minority-class samples from the PHEME and MediaEval datasets. The generated synthetic tweets undergo comprehensive multi-dimensional quality validation, including GPT-2 perplexity for fluency, VADER for sentiment alignment, LDA for thematic coherence, and MAUVE and Self-BLEU for semantic similarity and diversity, confirming high fidelity to original data. When incorporated into a DistilBERT-CNN hybrid classifier, the augmented datasets significantly improve performance, increasing F1-scores from 0.8767 to 0.9806 on PHEME and from 0.8074 to 0.9013 on MediaEval. Our approach substantially outperforms conventional augmentation techniques (Random Oversampling, Easy Data Augmentation, and Back Translation), with statistical significance validated through paired t-tests and large effect sizes. These findings establish LLM-based augmentation as a robust, scalable, and practical solution for addressing class imbalance in rumor and fake-news detection.

1 Introduction

The rapid proliferation of misinformation on social media, especially during crisis events, creates a complex and evolving challenge for public safety and effective communication. Rumors are pieces of information that have not been officially verified or refuted and

are typically shrouded in uncertainty [1]. In times of emergency, rumors can spread like wildfire, often exacerbating public anxiety and undermining trust in official sources [2]. As a result, robust rumor detection systems have become indispensable tools for identifying and curbing the spread of misinformation. These systems sift through vast amounts of social media data to pinpoint and flag dubious content, helping authorities and the public navigate conflicting narratives.

Despite their potential, one of the most significant challenges in building effective misinformation detec-

* Corresponding author.

Email address: Maskari@ardakan.ac.ir (M. Askarizade)

<https://dx.doi.org/10.22108/jcs.2026.148687.1200>

ISSN: 2322-4460



tion models lies in the inherent class imbalance present in real-world datasets. Minority-class examples, such as rumors or misused content, are often vastly outnumbered by majority-class instances, leading to skewed distributions that bias model training toward the majority class. Consequently, models trained on such imbalanced data tend to underperform on the under-represented target class, limiting their reliability and generalizability in practical applications.

To address this issue, we propose a novel data augmentation framework leveraging four state-of-the-art large language models (LLMs): DeepSeek-R1-Distill-Qwen-1.5B, Gemma-2-2B, GPT-4o-mini, and LLaMA-3.2-3B. Our methodology fine-tunes each model on the minority classes of two prominent rumor detection benchmarks, PHEME and MediaEval, generating high-quality synthetic tweets that enrich under-represented classes. The generated samples undergo a comprehensive multi-dimensional quality validation pipeline, including GPT-2 perplexity for fluency, VADER for sentiment alignment, LDA for thematic coherence, and MAUVE and Self-BLEU for semantic similarity and diversity. While our previous work [3] showed that a single generative model (GPT-2) can be used for rumor data augmentation, the present study significantly extends that line of research by systematically comparing multiple state-of-the-art LLMs, introducing a comprehensive multi-dimensional quality validation framework, and providing statistical, ablation, and cost analyses that were not part of the earlier study.

For classification, we employ a DistilBERT-CNN hybrid classifier, whose architectural choice is validated through comparisons against standard transformer-based models (BERT, DistilBERT, and RoBERTa). Our results show consistent and statistically significant improvements across all metrics compared to unaugmented baselines and conventional augmentation techniques (Random Oversampling, Easy Data Augmentation, and Back Translation). The improvements are validated through paired t-tests ($p < 0.05$) with large effect sizes ($Cohen's d > 0.8$). An ablation study confirms the critical importance of hashtag conditioning and calibrated temperature selection, while computational cost analysis demonstrates that open-weight models can be fine-tuned on a single GPU with negligible cost, ensuring practical accessibility.

The key contributions of this work are:

- (1) **Comparative Multi-LLM Augmentation Framework:** We propose a unified multi-LLM augmentation framework that leverages four architecturally diverse, instruction-tuned LLMs to generate synthetic, class-specific tweets for balancing minority classes in rumor detection datasets.

Each LLM is fine-tuned and evaluated independently under an identical protocol; no output-level fusion, weighting, or voting mechanism is applied. The framework is therefore best described as a comparative multi-LLM augmentation framework rather than an ensemble. Its value lies in the systematic isolation of model-specific effects and in the syntactic and semantic diversity that different LLMs contribute to the augmented corpus as a whole.

- (2) **Class-Conditional Fine-Tuning Strategy:** We introduce a class-conditional fine-tuning approach where each LLM is adapted to the specific linguistic and stylistic characteristics of the target minority class (rumor for PHEME, real for MediaEval). This instruction-based fine-tuning ensures that generated texts faithfully replicate authentic discourse patterns, outperforming zero-shot or prompt-only generation strategies.
- (3) **Comprehensive Quality Validation Framework:** We establish a rigorous validation pipeline that simultaneously evaluates synthetic data quality across multiple complementary dimensions: (i) linguistic fluency via GPT-2 perplexity, (ii) sentiment alignment via VADER, (iii) thematic coherence via LDA, (iv) semantic similarity via MAUVE, and (v) diversity via Self-BLEU. This multi-faceted assessment ensures that augmented data preserves the essential characteristics of the original corpus, a dimension often overlooked in previous augmentation studies.
- (4) **Exhaustive Comparison with Conventional Methods:** We systematically compare our LLM-based augmentation against established class imbalance techniques (Random Oversampling, Easy Data Augmentation, and Back Translation), demonstrating consistent superiority across both datasets. Statistical significance is validated through paired t-tests with large effect sizes.

The remainder of the paper is organized as follows. Section 2 reviews related work on rumor detection, data augmentation, and LLM-based generation, and also provides an overview of the four LLMs utilized in this study. Section 3 details our methodology, including preprocessing, class-conditional fine-tuning of multiple LLMs, and the controlled generation pipeline. Section 4 reports comprehensive evaluations, including quality assessment of synthetic text (perplexity, VADER, LDA, MAUVE, Self-BLEU), classification results, statistical significance analysis, ablation study, and computational cost analysis. Section 5 concludes and outlines future directions.



2 Related Works

The rapid proliferation of misinformation on social media platforms has motivated extensive research into automated rumor detection systems. Over the past decade, researchers have proposed a wide array of computational approaches to identify and mitigate the spread of unverified or false information, ranging from traditional feature-engineered classifiers to sophisticated deep learning architectures[4]. Despite these advances, several persistent challenges continue to hinder the development of robust and generalizable rumor detection models. Chief among these is the inherent class imbalance in real-world datasets, where verified rumors or fake news instances are substantially outnumbered by non-rumor content, leading to biased classifiers that prioritize majority-class predictions [4]. Furthermore, the evolving linguistic patterns, contextual nuances, and propagation dynamics of misinformation demand detection frameworks that are both adaptable and data-efficient. This section provides a comprehensive review of prior work, organized into six thematic parts: traditional machine learning approaches, deep learning-based methods, strategies for addressing class imbalance, a background on large language models utilized in this study, existing LLM-based augmentation frameworks, and a synthesis of the research gap that motivates our proposed approach.

2.1 Traditional Machine Learning Approaches

Early rumor detection research relied heavily on traditional machine learning with handcrafted features. Castillo et al. developed one of the first Twitter misinformation detectors by combining propagation patterns, user-credibility signals, and textual metadata with classifiers such as decision trees and SVM [5]. Kwon et al. extended this line by incorporating temporal, linguistic, and structural features, showing that careful feature engineering can substantially improve classification accuracy [6]. These studies established the importance of feature design, but their reliance on handcrafted attributes also limited adaptability to evolving misinformation patterns.

Subsequent work broadened the feature space and classifier families. Thakur et al. proposed a two-stage supervised machine-learning framework that first filters potential rumors and then analyzes their linguistic properties, achieving strong classification efficiency on large-scale Twitter data [7]. Wang et al. shifted the focus from rumor detection to rumor refutation, using machine-learning models, including XGBoost, to predict which users are likely to counteract disaster-related misinformation, highlighting the value of behavioral and interaction features for rumor management [8]. More recently, He et al. proposed a lifelong

machine-learning approach that accumulates knowledge across tasks and integrates BERT-based representations with an optimized genetic algorithm for feature selection, reporting robust performance even when training data are limited [9]. Overall, traditional machine-learning approaches provided interpretable feature-based pipelines and strong baselines, but their dependence on manual feature engineering motivated the subsequent shift toward neural and multimodal models.

2.2 Deep Learning-Based Approaches

Deep learning has significantly improved rumor detection by capturing complex patterns in textual, visual, and contextual data. Jing et al. introduced a Recurrent Neural Network (RNN) model that learns continuous representations of microblog events over time, outperforming traditional feature-based approaches in early rumor detection [10]. Similarly, CSI combines user behavior, article content, and propagation patterns using RNNs to improve fake news classification accuracy [11].

Beyond textual analysis, multimodal deep learning approaches have gained traction. A fusion model leveraging LSTM networks and an attention mechanism was proposed to integrate text, images, and social context for effective rumor classification [12]. Additionally, CNN-based models, such as Attention-CNN, incorporate time-series and sentiment polarity features to refine rumor classification [13]. Comparative studies highlight that ensemble methods, such as CNN + BiLSTM with attention, achieve state-of-the-art performance in fake news detection [14].

Hybrid deep learning models have also shown promising results in rumor detection. For example, a CNN-BiLSTM model incorporating GloVe embeddings has been proposed to enhance classification accuracy by capturing both spatial and sequential dependencies in textual data [15]. The model achieved 90.93% accuracy on a Kaggle dataset, outperforming baseline methods.

Additionally, Transformer-based models have demonstrated their effectiveness in mitigating rumor propagation. A BERT-based model was developed to detect false rumors solely from claim sentences before they go viral [16]. Further advancements in Transformer-based rumor detection integrate adversarial learning and graph-based methods to enhance model robustness [17].

Researchers have combined memory-augmented Transformers with Graph Convolutional Networks (GCNs) to better capture both local and global dependencies in rumor spread. The GCNs-MT frame-



work integrates multi-head attention mechanisms and LSTM cells to enhance sequential pattern recognition while leveraging graph-based structural information of rumor propagation [18]. Experiments on Chinese Weibo and curated English datasets demonstrated that this method significantly improves rumor classification performance, addressing key challenges posed by the rapid dissemination and complex nature of online misinformation.

These Transformer-based advancements highlight the potential of adversarial training, transfer learning, and structural propagation modeling in improving automatic rumor detection, offering promising solutions for combating misinformation on social media.

2.3 Addressing Class Imbalance in Rumor Detection

Class imbalance remains a critical challenge in rumor detection, often leading to biased models that favor the majority class. Several techniques have been proposed to mitigate this issue.

One widely used approach is Synthetic Minority Over-sampling Technique (SMOTE), introduced by Chawla et al., which generates synthetic samples for the minority class instead of simply duplicating existing instances [19]. SMOTE creates synthetic examples along the feature space between existing minority-class samples, improving model generalization and reducing overfitting. It has been widely applied in rumor detection tasks to enhance the detection rate of minority-class instances [16, 20].

Another strategy is cost-sensitive learning, where models assign higher misclassification penalties to the minority class to ensure better representation. A related study introduced an adaptive cost-sensitive loss function for stance classification, dynamically adjusting class weights during training [21]. Since stance detection plays a crucial role in rumor verification, improving the classification of rare stances like "query" and "deny" enhances early rumor detection. The proposed method automatically tunes the cost matrix, outperforming conventional approaches by increasing the F-score of rare classes. While focused on stance classification, this approach highlights the potential of adaptive cost-sensitive learning in improving imbalanced data handling for rumor detection.

Similarly, GERDA (Graph Embedding-based Rumor Data Augmentation) employs knowledge graph-based methods to simulate the generation process of rumors from a knowledge-driven perspective [22]. By incorporating graph embeddings and generative adversarial networks (GANs), GERDA generates diverse, high-quality rumor samples that help address data

imbalance and improve rumor classification performance. This method effectively models the evolution of misinformation, ensuring that the generated texts align with real-world rumor propagation patterns.

2.4 Large Language Models

LLMs have recently emerged as powerful generative models capable of producing coherent, contextually relevant, and semantically consistent text across a wide range of natural language processing tasks. Owing to their ability to capture complex linguistic patterns from large-scale corpora, LLMs have become increasingly popular for synthetic data generation, particularly in domains where annotated data are scarce or highly imbalanced. Unlike conventional over-sampling techniques that generate synthetic instances by interpolating feature representations, LLMs can produce diverse textual samples while preserving contextual meaning, making them especially suitable for data augmentation in misinformation and rumor detection. Recent studies have demonstrated that LLM-generated samples can improve the robustness and generalization capability of downstream classifiers by enriching the minority class with linguistically realistic examples.

To investigate the effectiveness of different generative architectures for rumor data augmentation, this study considers four representative state-of-the-art LLMs, namely DeepSeek-R1-Distill-Qwen-1.5B, Gemma-2-2B, GPT-4o-mini, and LLaMA-3.2-3B. These models represent diverse architectural designs, training paradigms, and reasoning capabilities, providing a comprehensive basis for evaluating the influence of LLM characteristics on synthetic rumor generation.

2.4.1 DeepSeek-R1-Distill-Qwen-1.5B

DeepSeek-R1-Distill-Qwen-1.5B is a reasoning-oriented large language model developed using a reinforcement learning-centric training paradigm [23]. The full model employs a Mixture-of-Experts (MoE) architecture with 671 billion parameters. In this study, we use the distilled open-weight variant DeepSeek-R1-Distill-Qwen-1.5B, which retains the reasoning capabilities of the original model while being small enough to fine-tune on a single GPU. This variant has approximately 1.5 billion parameters and supports long-context processing via Rotary Position Embedding (RoPE).

2.4.2 Gemma-2-2B

Gemma-2 is a family of lightweight decoder-only language models introduced by Google [24]. We use the Gemma-2-2B variant, which incorporates grouped-



query attention, interleaved local and global attention, and knowledge distillation. Its small size makes it suitable for scalable data augmentation in resource-constrained environments.

2.4.3 GPT-4o-mini

GPT-4o-mini is a compact, cost-efficient variant of the GPT-4 family, accessed via the OpenAI API. It combines large-scale autoregressive pre-training with Reinforcement Learning from Human Feedback (RLHF) to improve instruction following and contextual reasoning [25, 26]. We use GPT-4o-mini for supervised fine-tuning and synthetic data generation through the OpenAI platform.

2.4.4 LLaMA-3.2-3B

LLaMA-3.2 is an open-weight multilingual large language model developed by Meta AI [27]. We use the LLaMA-3.2-3B text-only variant, which is trained on approximately nine trillion multilingual tokens and supports context windows of up to 128k tokens. It employs grouped-query attention to improve inference efficiency.

The diversity among these representative LLMs in terms of architecture, training methodology, reasoning capability, and computational complexity provides a valuable foundation for investigating how different generative models influence the quality of synthetic data. Their complementary characteristics have motivated their increasing adoption in data augmentation tasks across misinformation detection, fake news classification, and other low-resource natural language processing applications.

2.5 LLM-Based Data Augmentation

The emergence of LLMs has fundamentally changed the landscape of data augmentation for rumor detection. Unlike traditional oversampling techniques, which merely duplicate or interpolate existing feature representations, LLMs are capable of generating semantically coherent and contextually relevant textual samples that closely resemble naturally occurring misinformation. Consequently, recent research has increasingly adopted LLM-based generation to address the persistent challenge of class imbalance while simultaneously enriching the linguistic diversity of training corpora.

One of the earliest efforts in this direction is RumorLLM proposed by Lai et al. [28], which fine-tunes a language model specifically on rumor-related writing styles to generate synthetic fake-news samples. Their experiments demonstrated that LLM-generated data can substantially improve downstream classifi-

cation performance, achieving notable gains in F1-score and AUC-ROC on benchmark datasets such as BuzzFeed and PolitiFact. Furthermore, an innovative transformer-based generative adversarial approach has been proposed to improve rumor detection [17]. In this approach, a GAN-style model is employed where a generator produces uncertain or conflicting voices that further polarize conversation threads. This process pressures the discriminator to learn more robust, rumor-indicative features from augmented, challenging examples. Leveraging the strong natural language generation capabilities of transformers, this method captures low-frequency yet salient discriminative patterns, ultimately outperforming state-of-the-art methods across multiple benchmark datasets. Similarly, Askarizade [3] employed GPT-2 to generate rumor-like tweets for balancing minority classes, confirming that generative language models can effectively mitigate data scarcity in rumor detection tasks. These studies established the feasibility of generative data augmentation; however, they relied on a single language model and therefore could not investigate how different LLM architectures influence the quality and usefulness of synthetic data.

The present work directly extends this line of research in several fundamental ways. First, we systematically compare four state-of-the-art instruction-tuned LLMs under an identical class-conditional fine-tuning and generation protocol. This controlled comparison allows us to isolate the effects of model architecture, scale, and training methodology on synthetic data quality and downstream classification, which was not possible with a single-model study. Second, we introduce a comprehensive multi-dimensional quality validation pipeline that evaluates fluency (GPT-2 perplexity), sentiment alignment (VADER), thematic coherence (LDA), semantic similarity (MAUVE), and diversity (Self-BLEU). This goes far beyond the evaluation in [3] and provides a holistic view of synthetic data quality. Third, we provide statistical significance testing, an ablation study on hashtag conditioning and temperature, and a detailed computational cost analysis. While fine-tuning LLMs for augmentation is not new, and using multiple LLMs is not a contribution in itself, our work provides the first systematic, reproducible, and practically oriented framework that quantifies how different LLMs and design choices affect augmentation effectiveness in rumor detection. These extensions transform a proof-of-concept into a rigorous investigation that yields actionable guidelines for model selection, quality assessment, and deployment.

Recent studies have expanded this paradigm by exploiting more advanced instruction-tuned LLMs. Chalehchaleh et al. [29] investigated multilingual fake-news detection using Llama 3 under zero-shot and few-



shot learning scenarios. Their results demonstrated that selectively augmenting the minority class improves classification performance across multiple languages. Likewise, Tong et al. [30] introduced the Generate First, Then Sample framework, which combines synthetic text generation with reinforcement learning to determine the optimal ratio between real and generated samples. These approaches illustrate the growing interest in leveraging LLMs for intelligent data augmentation. Nevertheless, they primarily depend on prompt engineering rather than task-specific fine-tuning, making it more difficult to preserve the stylistic, semantic, and contextual characteristics of authentic rumor content.

Beyond synthetic data generation, several recent studies have explored alternative applications of LLMs for misinformation detection. For example, the Shortcut Mitigation Framework (SMF) proposed by Wan et al. [31] utilizes LLM-based paraphrasing, summarization, and sentiment normalization to improve model robustness by reducing shortcut learning. Similarly, the Topic-Aware Chain-of-Reasoning (TA-CoR) framework [32] integrates topic clustering with LLM reasoning to improve explainable rumor detection, while recent methods such as LLM-Enhanced Virtual Node Detection [33] and SharP [34] employ structural augmentation and prompt-guided learning strategies to enhance graph-based and transformer-based classifiers. Although these methods demonstrate the versatility of LLMs in misinformation detection, their primary objective is improving reasoning ability, explainability, or representation learning rather than directly addressing class imbalance through minority-class data generation.

Table 1 summarizes representative LLM-based approaches for misinformation detection and highlights their methodological characteristics. Although these studies have substantially advanced the field, most rely on a single language model, employ prompt-based generation without task-specific fine-tuning, and evaluate the generated data primarily through downstream classification performance. Comprehensive comparisons among multiple state-of-the-art LLMs and systematic evaluations of synthetic data quality remain relatively limited.

The comparison reveals several important research gaps. First, existing studies predominantly employ a single LLM, making it difficult to determine whether the observed improvements stem from the augmentation strategy itself or from the characteristics of a particular model. Second, most approaches rely on prompt engineering alone and do not incorporate task-specific, class-conditional fine-tuning, which is essential for generating samples that preserve the linguistic

style and semantic characteristics of rumor-related content. Third, the evaluation of synthetic data is generally limited to downstream classification performance, while comprehensive assessments of text quality, including fluency, semantic consistency, sentiment preservation, and topical coherence, are rarely conducted. Finally, systematic comparisons of multiple state-of-the-art LLMs under a unified experimental protocol remain largely unexplored.

Motivated by these limitations, this study proposes a unified framework for systematically evaluating multi-LLM data augmentation in rumor detection. Four state-of-the-art instruction-tuned language models are fine-tuned using an identical class-conditional training strategy and evaluated under the same experimental settings. Rather than claiming multi-LLM usage or fine-tuning as novelties in themselves, we position our contribution as the systematic investigation and comprehensive evaluation that this unified protocol enables. Specifically, our framework compares the effectiveness of multiple LLMs for synthetic data generation and performs a multi-dimensional assessment of the generated samples using complementary measures of linguistic fluency, sentiment preservation, topical consistency, semantic similarity, diversity, and downstream classification performance. This unified evaluation provides a more comprehensive understanding of the capabilities of modern LLMs for addressing class imbalance in rumor detection while offering practical guidance for selecting appropriate generative models in misinformation analysis.

3 Methodology

We perform data augmentation using LLMs to generate class-specific synthetic tweets that enhance the training datasets for rumor detection. The fine-tuned LLMs are trained on two prominent rumor datasets: PHEME and MediaEval, each of which has an imbalanced class distribution. In the PHEME dataset, rumor tweets represent the minority class, while in MediaEval, real (non-rumor) tweets are underrepresented. Our goal is to augment these minority classes by generating high-quality, class-specific synthetic tweets using fine-tuned LLMs, thereby improving the downstream classification models' performance.

The methodology consists of several key stages, as illustrated in Figure 1, ensuring the generated text maintains fluency, contextual integrity, and alignment with real-world discourse. We begin by extracting all minority-class samples, rumor tweets for PHEME and real tweets for MediaEval. These subsets serve as the basis for fine-tuning the LLMs, enabling them to model the linguistic and contextual characteristics of



Table 1. Comparison of Representative LLM-based Data Augmentation Methods for Misinformation Detection.

Study	Year	LLM(s) Used	Fine-Tuned	Multi-LLM Comparison	Quality Evaluation	Primary Objective
RumorLLM [28]	2024	RumorLLM (fine-tuned)	✓	×	Limited	Rumor data augmentation
Askarizade [3]	2025	GPT-2	✓	×	Perplexity, Sentiment, Topic	Rumor data augmentation
Chalehchaleh <i>et al.</i> [29]	2025	Llama 3	×	×	Limited	Multilingual fake-news augmentation
Tong <i>et al.</i> [30]	2025	ChatGLM-6B	×	×	Limited	Fake-news augmentation
SMF [31]	2025	MISTRALV3, LLAMA3	×	×	No	Shortcut mitigation
TA-CoR [32]	2026	GPT-3.5-turbo	✓	×	No	Explainable rumor detection
LLM-Enhanced Virtual Node [33]	2026	Qwen-Plu, DeepSeek	×	×	No	Structural augmentation
SharP [34]	2026	Zephyr-7b-beta	×	×	Limited	Prompt-guided augmentation
Proposed Framework	2026	DeepSeek-R1-Distill-Qwen-1.5B, Gemma-2-2B, GPT-4o-mini, LLaMA-3.2-3B	✓	✓	Perplexity, VADER, LDA, MAUVE, Self-BLEU	Systematic multi-LLM comparison and class-conditional augmentation

^a The original study does not report a specific underlying LLM or adopts a model-independent approach.

the underrepresented class.

Once the minority-class texts are isolated, each pre-trained LLM is fine-tuned on its respective subset. This step allows the model to capture the stylistic, syntactic, and semantic patterns characteristic of the targeted class, such as the propagation structure of rumors or the factual tone of real tweets. After fine-tuning, the LLM generates synthetic samples, which are then incorporated into the original training dataset to balance class distributions. This augmentation not only mitigates class imbalance but also enriches the training data with varied, yet realistic, examples, improving the robustness and generalizability of the classification model.

The complete augmentation pipeline is depicted in Figure 1, which illustrates the process from minority-class extraction through LLM fine-tuning to synthetic generation and dataset augmentation. The resulting augmented datasets are subsequently used to train a DistilBERT-CNN hybrid classifier. A detailed evaluation of both the quality of the generated synthetic data and the downstream classification performance is presented in Section 4.

Figure 1 visually represents this process, illustrating

how the dataset is augmented and evaluated in a structured pipeline. By leveraging LLM-based augmentation, we aim to improve rumor classification while maintaining data quality and fluency.

3.1 Preprocessing

Before applying data augmentation, we preprocessed the dataset to enhance the quality of textual data. The preprocessing steps include:

- **Lowercasing:** All texts are converted to lowercase to ensure uniformity.
- **Removing URLs:** Any hyperlinks, including "http://", "https://", and "www." links, are eliminated.
- **Tokenization:** Text is split into individual words using tokenization techniques.
- **Stop Word Removal:** Common stop words that do not contribute to meaningful context are filtered out.
- **Punctuation Removal:** Punctuation marks are removed to focus on meaningful words.

This preprocessing step ensures cleaner and more structured input for both augmentation and classification.



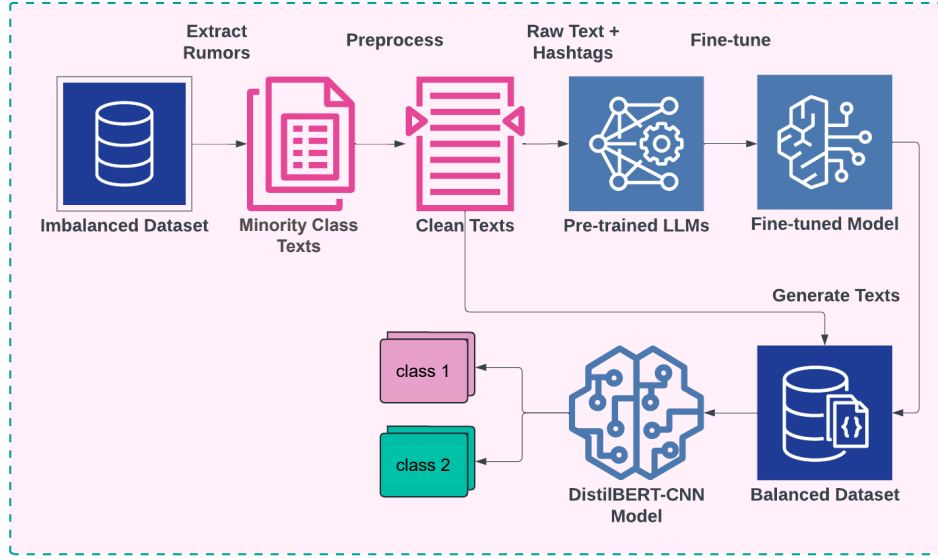


Figure 1. Workflow for Balancing Rumor Detection Dataset Using LLM-Based Synthetic Text Generation and DistilBERT-CNN Classification.

3.2 Data Augmentation using Large Language Models

LLMs have demonstrated strong capabilities in understanding and generating human-like text across a wide range of domains. In this study, we leverage the generative capacity of several state-of-the-art instruction-tuned LLMs to generate high-quality synthetic data tailored to the minority class in two benchmark rumor detection datasets: *PHEME* and *MediaEval*. The primary goal is to mitigate class imbalance by augmenting underrepresented classes with context-aware and semantically coherent synthetic tweets.

3.2.1 Prompt Engineering Strategy

To ensure consistent and reproducible data augmentation across all language models, a unified prompt engineering strategy was adopted. The same prompt template was used for all four LLMs, allowing differences in the generated samples to be attributed solely to the characteristics of the underlying models rather than variations in prompting.

The prompts were designed following three principles: (1) explicit specification of the target class, (2) preservation of the original discussion topic through hashtags, and (3) generation of concise Twitter-style posts that resemble authentic user-generated content. During fine-tuning, each training instance was transformed into an instruction–response pair, where the instruction specifies the target class and topic, and the response corresponds to the original tweet. Table 2 presents the prompt templates used in this study.

Table 2. Prompt Templates Used for Class-conditional Fine-tuning and Data Generation.

Target Class	Prompt Template
Rumor	Generate a realistic Twitter post that expresses a rumor about the topic {hashtag}. The tweet should resemble natural user-written content, preserve the discussion topic, and contain only the generated tweet.
Real	Generate a realistic Twitter post that expresses a real (non-rumor) statement about the topic {hashtag}. The tweet should resemble authentic Twitter content, preserve the discussion topic, and contain only the generated tweet.

The placeholder {hashtag} was replaced with the hashtag extracted from each original tweet, enabling the generated samples to remain topically aligned with the corresponding discussion thread while encouraging linguistic diversity.

3.2.2 Fine-Tuning LLMs

Following prompt construction, the instruction–response pairs were used to fine-tune each LLM for class-conditional text generation. Let the resulting training dataset be:

$$\mathcal{D} = \left\{ (x^{(i)}, y^{(i)}) \right\}_{i=1}^N \quad (1)$$

where:

- $x^{(i)}$ is the class-specific prompt,



- $y^{(i)} = (y_1^{(i)}, y_2^{(i)}, \dots, y_T^{(i)})$ is the tweet token sequence.

Fine-tuning aims to adapt the model's parameters θ so that, conditioned on $x^{(i)}$, the model can generate a fluent and class-representative tweet $y^{(i)}$. The training objective minimizes the conditional negative log-likelihood:

$$\mathcal{L}_{\text{fine-tune}}(\theta) = - \sum_{i=1}^N \sum_{t=1}^T \log P_{\theta} \left(y_t^{(i)} \mid y_{<t}^{(i)}, x^{(i)} \right) \quad (2)$$

where:

- $P_{\theta}(\cdot)$ is the model's predicted probability distribution,
- $y_{<t}^{(i)}$ represents the preceding tokens in the sequence,
- $x^{(i)}$ provides the instructional conditioning.

Fine-tuning was performed separately for each dataset according to its minority class. For the PHEME dataset, the models were fine-tuned to generate *rumor* tweets, whereas for the MediaEval dataset they were fine-tuned to generate *real* tweets. The same prompt templates, fine-tuning procedure, and hyperparameter configuration were employed for all four LLMs to ensure a fair comparison of their augmentation capabilities.

3.2.3 Controlled and Context-Aware Text Generation

Once fine-tuned, the models are employed to generate synthetic tweets that reflect the characteristics of the minority class. The generation process incorporates several key strategies to ensure quality and relevance:

- **Controlled Generation via Prompt Engineering:** Prompts explicitly guide the model toward the minority class. Examples include “Generate a real news tweet” or “Write a tweet that contains a rumor,” ensuring that the generated output adheres to the desired label semantics.
- **Hashtag Conditioning for Topical Coherence:** To enhance the topical relevance of the generated tweets, we apply hashtag-based conditioning during generation. Specifically, for each tweet selected from the minority class, its associated hashtag (if present) is extracted and incorporated into the generation prompt. The prompt is then reformulated to reflect both the class label and topical focus, for example: “Generate a real news tweet about {hashtag}” for the MediaEval dataset or “Write a tweet that contains a rumor about {hashtag}” for the PHEME dataset. This strategy ensures that the output remains not only class-consistent but also contextually grounded

in relevant online discourse.

- **Temperature Scaling for Diversity Control:** The generation process leverages temperature scaling to control the diversity and creativity of the generated outputs. A higher temperature encourages more varied and exploratory outputs, while a lower temperature promotes deterministic and conservative responses. For each of the four models, a specific temperature value was selected to balance relevance and diversity in generation. The exact values and rationale for each model's temperature setting are reported and discussed in Section 4.3.

The proposed data augmentation strategy offers several theoretical advantages for enhancing rumor detection performance. First, it addresses class imbalance by generating additional training samples specifically for the minority class, thereby improving model robustness and fairness in classification. Second, it promotes better generalization by exposing the classifier to a broader and more diverse set of examples, which go beyond the limitations of the original dataset. Third, by employing instruction-based prompts combined with hashtag grounding, the generated samples maintain contextual relevance and align closely with real-world discourse patterns and topical trends. Moreover, by comparing four generative language models with distinct architectures, scales, and training priors, the framework exposes a broader range of syntactic and semantic structures than any single model could provide. We emphasize that this is a comparative multi-LLM augmentation framework, not a model-level ensemble: each LLM is fine-tuned and evaluated independently, and no aggregation, weighting, voting, or fusion of model outputs is performed. The diversity benefit arises because different models characterize the minority class differently, which allows practitioners to either select the best-performing model for a given dataset or, optionally, pool synthetic samples from multiple models into a single augmented corpus.

3.3 Classification Method: DistilBERT-CNN Hybrid Model

In this classification approach, our primary objective is to compare the effectiveness of different LLMs for data augmentation in the context of rumor detection. To ensure a fair and consistent evaluation across augmentation strategies, we adopt a single, robust classification model that performs well across datasets. Specifically, we integrate DistilBERT with a CNN, forming a hybrid model named DistilBERT-CNN. This model combines the rich contextual embeddings of DistilBERT with the local feature extraction capabilities of CNN layers, effectively capturing both global and local semantic patterns in text. Figure 2 illustrates the



overall architecture of the proposed DistilBERT-CNN model.

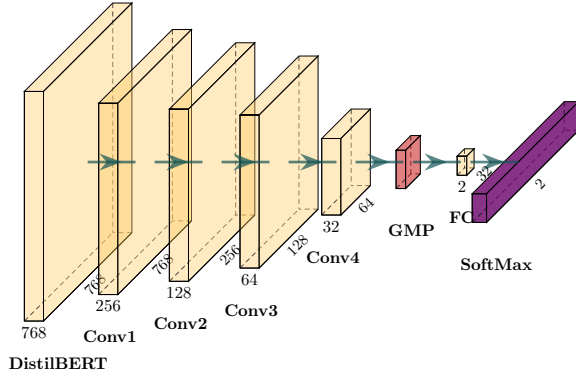


Figure 2. The Architecture of the DistilBERT-CNN Hybrid Model Used for Rumor Classification.

DistilBERT is a lighter and faster version of BERT that retains approximately 97% of BERT’s performance while being 60% smaller and running twice as fast [35]. It is trained using knowledge distillation, where a larger pre-trained BERT model serves as a teacher, and the DistilBERT model learns a compressed yet effective representation of language. By removing certain components such as the next sentence prediction task and reducing the number of attention layers, DistilBERT maintains strong contextual understanding while improving computational efficiency. This makes it an ideal choice for real-world NLP applications, including rumor classification.

As shown in Figure 2, the DistilBERT-CNN model consists of two main components: a DistilBERT encoder for extracting contextualized representations from input text and CNN layers for refining these embeddings. The classification process begins with tokenizing the input text, converting it into token IDs and attention masks, and feeding these inputs into DistilBERT. The model generates contextual embeddings, where each token is represented by a 768-dimensional hidden vector, capturing deep semantic information.

To leverage CNN for feature extraction, the embeddings are transposed to align with 1D convolutional layers. The model applies a series of four convolutional layers, each followed by batch normalization and a ReLU activation function. These layers progressively reduce the feature map dimensions from 768, 256, 128, 64, 32, helping the network learn essential patterns while maintaining computational efficiency. The batch normalization layers stabilize training and enhance generalization by normalizing intermediate feature distributions.

Once the convolutional layers extract meaningful patterns, a global max pooling layer is applied to retain the most significant features from each feature

map, reducing the output to a compact feature representation. To prevent overfitting, a dropout layer of 0.3 is introduced before the final fully connected layer, which maps the extracted features to the output classes, distinguishing between rumor and non-rumor tweets. The final output consists of raw logits, which are passed through a softmax function to generate classification probabilities.

For training, the model utilizes Cross-Entropy Loss, which quantifies the difference between predicted and actual labels. The Adam optimizer is employed to dynamically adjust learning rates, ensuring efficient and stable convergence. By integrating DistilBERT’s contextual representations with CNN’s ability to capture local dependencies, this hybrid model effectively balances computational efficiency and representation learning. The combination of these architectures enables the model to capture both global semantic meaning and local structural patterns, making it well-suited for rumor classification tasks.

In summary, our methodology systematically addresses class imbalance in rumor detection datasets by leveraging fine-tuned LLMs to generate high-quality, class-specific synthetic tweets. Through careful preprocessing, controlled text generation with multiple state-of-the-art LLMs, and robust augmentation strategies, we enrich minority classes with diverse and realistic synthetic samples. The augmented datasets are subsequently used to train a DistilBERT-CNN hybrid classifier, and a comprehensive evaluation of both synthetic data quality and classification performance is presented in Section 4.

4 Evaluation and Experiments

In this section, we evaluate the quality and effectiveness of our LLM-based augmentation approach using a comprehensive multi-faceted evaluation pipeline. We begin by describing the PHEME and MediaEval datasets as our benchmarks. Our evaluation is organized into four complementary parts.

First, we assess classification performance using standard metrics (Accuracy, Precision, Recall, and F1-score) on both the original and LLM-augmented datasets using a DistilBERT-CNN hybrid classifier. This evaluation quantifies the practical impact of augmentation on downstream rumor detection. We further validate our architectural choice by comparing DistilBERT-CNN against standard transformer-based classifiers (BERT, DistilBERT, and RoBERTa) to ensure that observed improvements are attributable to augmentation rather than classifier selection. Statistical significance of the performance gains is assessed through paired t-tests with effect size analysis (Co-



hen's d).

Second, we conduct a multi-dimensional quality assessment of the generated synthetic tweets to ensure they preserve essential linguistic, semantic, and thematic characteristics of the original data:

- **Fluency Assessment:** We evaluate linguistic naturalness and coherence using GPT-2 perplexity. Lower perplexity scores indicate more fluent and structurally plausible outputs.
- **Semantic Alignment:** We measure distributional similarity between generated and original texts using MAUVE, which computes the divergence frontier between two sets of embeddings in a pretrained language model space.
- **Diversity Assessment:** We quantify lexical and syntactic diversity using Self-BLEU, where lower scores indicate higher diversity and reduced mode collapse.
- **Sentiment Alignment:** We apply VADER sentiment analysis to verify that the emotional tone distribution in augmented data mirrors that of the source corpus, ensuring sentiment realism and preventing emotional bias.
- **Thematic Coherence:** We employ Latent Dirichlet Allocation (LDA) topic modeling, with the number of topics selected via coherence score optimization, to confirm that the latent thematic structure of the original data is preserved after augmentation.

Third, we conduct an ablation study to dissect the contribution of our key design choices, including hashtag conditioning and temperature calibration, using our best-performing model (GPT-4o-mini) on the PHEME dataset.

Fourth, we provide a computational cost and scalability analysis, reporting fine-tuning times, inference times, memory requirements, and estimated costs for all four LLMs. This analysis demonstrates the practical feasibility of our approach across diverse resource settings, including the accessibility of open-weight models on a single GPU with negligible cost.

This comprehensive evaluation strategy ensures both the practical utility (classification performance) and the intrinsic quality (fluency, semantic alignment, diversity, sentiment, and topics) of our augmentation framework, while also validating our design choices and confirming practical applicability.

4.1 Datasets

We conduct our experiments on two benchmark datasets commonly used in rumor and fake content detection tasks. In selecting these datasets, our focus was on real-world challenges: specifically, datasets

that are imbalanced and contain a large number of tweets, as these better reflect practical deployment conditions in misinformation detection systems.

While our primary focus is on rumor and misinformation detection, we also include the MediaEval dataset, which is primarily designed for fake news detection. Despite this difference, MediaEval exhibits significant class imbalance and provides rich textual data, making it suitable for evaluating the robustness and generality of our proposed approach across related tasks. By incorporating MediaEval alongside a dedicated rumor detection dataset, we aim to demonstrate that our method generalizes well to different yet relevant domains involving false or misleading information.

- **PHEME** [36] is a well-known dataset for rumor detection containing Twitter conversations collected from five major breaking news events: Charlie Hebdo, Ferguson, Germanwings Crash, Ottawa Shooting, and Sydney Siege. Each conversation includes a source tweet, replies, and associated metadata. The dataset contains 5,802 annotated tweets, with 1,972 labelled as *rumours* and 3,830 as *non-rumours*. The dataset includes a predefined test set of 643 tweets, consisting of 226 rumours and 417 non-rumours, which we use without modification.
- **MediaEval 2015** [37] is a fake news dataset consisting of tweets that include images, with binary labels: *real* (accurate use of image) or *fake* (misused or manipulated images). We focus on the tweet text and its label, ignoring visual features. The training set includes 5,008 real and 6,840 fake tweets, while the test set includes 1,217 real and 2,564 fake tweets. Similar to PHEME, the dataset exhibits class imbalance, with real tweets forming the minority class.

Table 3. Statistics of PHEME and MediaEval Datasets Used in Our Study.

Dataset	Split	Positive Class	# Positive	# Negative
PHEME	Train	Rumour	1,746	3,413
	Test	Rumour	226	417
MediaEval	Train	Real	5,008	6,840
	Test	Real	1,217	2,564

In both datasets, we identify the minority class, *rumour* in PHEME and *real* in MediaEval, as the target for generative data augmentation.

4.2 Performance Evaluation

We assess classification performance using four standard metrics, Accuracy, Precision, Recall, and F1-



score, which are particularly informative in the context of imbalanced datasets such as PHEME and MediaEval, where the minority class (rumours in PHEME, real tweets in MediaEval) is of primary interest.

Accuracy: Accuracy measures the overall proportion of correctly classified instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

where:

- **TP (True Positives):** Minority-class tweets correctly predicted (rumours in PHEME, real tweets in MediaEval).
- **TN (True Negatives):** Majority-class tweets correctly predicted (non-rumours in PHEME, fake tweets in MediaEval).
- **FP (False Positives):** Majority-class tweets incorrectly predicted as minority.
- **FN (False Negatives):** Minority-class tweets incorrectly predicted as majority.

Although accuracy provides an overall correctness measure, it can be misleading when classes are highly imbalanced.

Precision: Precision quantifies the proportion of predicted minority-class tweets that are actually from the minority class:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

Recall: Recall (also known as sensitivity) measures the proportion of true minority-class tweets that are correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

F1-score: The F1-score is the harmonic mean of Precision and Recall, providing a balanced measure when class distributions are skewed:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

By focusing on these metrics, especially Precision, Recall, and F1-score for the minority class, we obtain a comprehensive picture of our model's ability to detect rumours in PHEME and real tweets in MediaEval.

4.3 Classification Evaluation

This subsection presents the experimental setup and results for evaluating the effectiveness of our proposed LLM-based data augmentation framework in improving rumor detection performance. We first describe the fine-tuning configuration employed to adapt each of the four large language models to the linguistic and

stylistic characteristics of the target minority class. Next, we detail the cross-validation evaluation protocol used to ensure a rigorous and unbiased assessment of the augmented datasets. Following this, we present a comparative analysis of the classification performance achieved by each LLM-based augmentation strategy, highlighting the relative strengths of different generative architectures. Finally, we compare our proposed approach against three conventional class imbalance handling techniques: Random Oversampling, Easy Data Augmentation, and Back Translation to demonstrate the superiority of LLM-powered augmentation in enhancing rumor and fake-news detection. All experiments were conducted under identical training and evaluation conditions to ensure a fair and comprehensive comparison.

4.3.1 LLM Fine-Tuning Configuration

To generate high-quality synthetic data for rumor detection, we fine-tuned four state-of-the-art large language models: DeepSeek-R1-Distill-Qwen-1.5B, Gemma-2-2B, GPT-4o-mini, and LLaMA-3.2-3B. The fine-tuning process was designed to adapt each model to the specific linguistic and stylistic characteristics of the target minority class (rumor for PHEME, real for MediaEval) through class-conditional instruction-based training.

For the open-weight models (DeepSeek-R1-Distill-Qwen-1.5B, Gemma-2-2B, and LLaMA-3.2-3B), we employed a unified fine-tuning pipeline implemented using the Unsloth framework combined with Parameter-Efficient Fine-Tuning (PEFT) via Low-Rank Adaptation (LoRA). This approach significantly reduces the number of trainable parameters while preserving the generative quality of the full model. Each model was fine-tuned with 4-bit NF4 quantization to reduce memory footprint, and LoRA was applied to the query, key, value, output, gate, up, and down projection modules with a rank of $r = 16$ and scaling factor $\alpha = 16$. Training was conducted for 50 optimization steps using the AdamW optimizer with an 8-bit implementation, a learning rate of 2×10^{-4} , a linear scheduler with 5 warm-up steps, and weight decay of 0.01. Gradient checkpointing was enabled, and a per-device batch size of 2 with 4 gradient accumulation steps was used. All experiments were executed on an NVIDIA Tesla P100 GPU.

For GPT-4o-mini, fine-tuning was performed using the OpenAI supervised fine-tuning platform. The training data was formatted as instruction-response pairs matching the same prompt templates used for the open-weight models. The learning rate multiplier was set to 1.8, and the batch size was configured as 4 for PHEME and 12 for MediaEval, consistent with



OpenAI’s recommendations. The model was trained for 50 epochs with distinct random seeds (1064277297 for PHEME, 287327811 for MediaEval) to ensure reproducibility. All fine-tuning details are summarized in Table 4.

For downstream rumor classification, we adopted a stratified 5-fold cross-validation protocol implemented in PyTorch. In each fold, the original dataset was first partitioned into training (80%) and validation (20%) subsets while preserving the class distribution. LLM fine-tuning and synthetic data generation were performed exclusively using the training partition, and the generated samples were added only to the corresponding training set to mitigate class imbalance. The validation partition remained completely unchanged and consisted solely of original tweets throughout the evaluation process, ensuring a strict separation between training and evaluation data with no data leakage.

The DistilBERT-CNN hybrid classifier was trained for 50 epochs on the augmented training set using the Adam optimizer with a learning rate of 2×10^{-5} and cross-entropy loss. Classification performance was evaluated on the original validation set of each fold using Accuracy, Precision, Recall, and F1-score, and the final results were obtained by averaging the metrics across all five folds.

4.3.2 Comparison of LLM-Based Augmentation Strategies

Following fine-tuning, each model was used to generate synthetic tweets for the minority class of each dataset. Text generation employed stochastic sampling, where the temperature parameter was adjusted individually for each LLM to balance diversity and semantic coherence. The temperature values were determined through a preliminary pilot study in which multiple candidate values (0.1–0.8) were evaluated for each model. The generated samples were assessed through both qualitative inspection (fluency, contextual relevance, and repetition) and the quantitative quality metrics adopted in this work, including perplexity, sentiment consistency, and topic consistency. The final temperature for each LLM was selected as the best trade-off between generating diverse outputs and preserving realistic, contextually coherent tweets. The selected temperature values are summarized in Table 5. During generation, the maximum number of generated tokens was limited to 64, while nucleus sampling ($top-p = 0.9$) and a repetition penalty of 1.1 were applied consistently across all open-weight models.

For the MediaEval dataset, the LLMs collectively generated 1,832 synthetic tweets labeled as “real,” while for the PHEME dataset, 1,621 synthetic tweets

labeled as “rumor” were produced. The examples below, generated by GPT-4o-mini, illustrate the high quality and contextual relevance of the synthetic outputs for each dataset:

- **MediaEval (Real-like Tweets):**

- “RT : A mob of marathoners heading over to Staten Island. No race They’re going to help out instead.”
- “Amazing cover from New York Magazine. A view of Manhattan from above when Sandy hit”

- **PHEME (Rumor-like Tweets):**

- “Ferguson police chief says he’s received death threats, claims whites are being killed by blacks in the community.”
- “Ottawa police report that a man is accused of shooting at least one of the hostages currently being held in the city.”

Table 6 summarizes the classification performance of the proposed DistilBERT-CNN model on the PHEME and MediaEval datasets before and after LLM-based data augmentation. Since all evaluation metrics were computed exclusively on the untouched original validation data, the reported improvements reflect the contribution of the synthetic samples to classifier training rather than the presence of augmented data during testing.

For both datasets, the baseline performance without augmentation demonstrates strong results, particularly on the PHEME dataset (F1-score: 0.8767). However, the addition of synthetic data generated by LLMs leads to a marked improvement in classification outcomes across all metrics. The enhancement is especially significant for the MediaEval dataset, which initially had a lower F1-score of 0.8074. Despite the short length of tweets and the prevalence of URLs in MediaEval, which often limit semantic richness, LLM-generated data still managed to improve the model’s generalization ability.

Among the augmentation strategies, GPT-4o-mini consistently achieves the highest scores on both datasets (F1-score: 0.9806 for PHEME, 0.9013 for MediaEval). Models such as DeepSeek-R1-Distill-Qwen-1.5B and LLaMA-3.2-3B also deliver robust performance, indicating their effectiveness in generating label-consistent and contextually relevant samples that benefit the classification task. Interestingly, even models that are comparatively less complex, such as Gemma-2-2B, still offer measurable gains over the baseline. This suggests that the added data diversity and volume alone, even if generated by smaller models, contribute positively to model training.



Table 4. Fine-tuning Configuration of the LLMs.

Parameter	DeepSeek-R1-Distill-Qwen-1.5B, Gemma-2-2B, LLaMA-3.2-3B	GPT-4o-minio-mini
Fine-tuning framework	Unsloth + PEFT	OpenAI Fine-Tuning API
Training method	LoRA	Supervised Fine-Tuning
Quantization	4-bit NF4	Managed by OpenAI
Maximum sequence length	2048	OpenAI default
LoRA rank (r)	16	—
LoRA alpha	16	—
LoRA dropout	0	—
Target modules	q-proj, k-proj, v-proj, o-proj, gate-proj, up-proj, down-proj	—
Gradient checkpointing	Enabled	—
Optimizer	AdamW 8-bit	Managed by OpenAI
Learning rate	2×10^{-4}	LR multiplier = 1.8
Scheduler	Linear	Managed by OpenAI
Weight decay	0.01	Managed by OpenAI
Per-device batch size	2	4 (PHEME), 12 (MediaEval)
Gradient accumulation	4	—
Warm-up steps	5	—
Random seed	3407	1064277297 (PHEME), 287327811 (MediaEval)
Training duration	50 optimization steps	50 epochs

Table 5. Temperature Settings for Synthetic Data Generation.

Model	Temperature
DeepSeek-R1-Distill-Qwen-1.5B	0.1
Gemma-2-2B	0.5
GPT-4o-mini	0.4
LLaMA-3.2-3B	0.1

4.3.3 Statistical Significance Analysis

To determine whether the observed differences in classification performance are statistically significant, we performed paired t-tests comparing the best-performing model (GPT-4o-mini) against each other augmentation method across the five cross-validation folds. This analysis ensures that the reported improvements are not attributable to random variation in the evaluation protocol.

Table 7 reports the detailed statistical results, including the mean F1-score difference, p-values, and effect sizes (Cohen’s d). On PHEME, GPT-4o-mini significantly outperforms all other methods ($p < 0.05$ for all comparisons), with effect sizes ranging from 1.63

to 14.95, indicating very large practical significance. On MediaEval, the improvement over DeepSeek-R1-Distill-Qwen-1.5B (0.9040 vs. 0.8860, diff = 0.0180) is statistically significant ($p = 0.0015$, Cohen’s $d = 3.47$), confirming that the performance gain is not due to random variation. All comparisons yield large to very large effect sizes ($d > 0.8$), demonstrating the substantial practical impact of LLM-based augmentation across all configurations.

The consistently large effect sizes across all comparisons underscore that the performance improvements are not merely statistically significant but also practically meaningful. The largest effect size is observed for the unaugmented baseline on both datasets ($d = 14.95$ for PHEME, $d = 11.26$ for MediaEval), confirming that LLM-based augmentation provides substantial and robust gains over no augmentation. Even for the closest competitor (DeepSeek-R1-Distill-Qwen-1.5B), the effect size remains large ($d = 1.63$ on PHEME, $d = 3.47$ on MediaEval), indicating that GPT-4o-mini’s superiority is both statistically and practically significant.

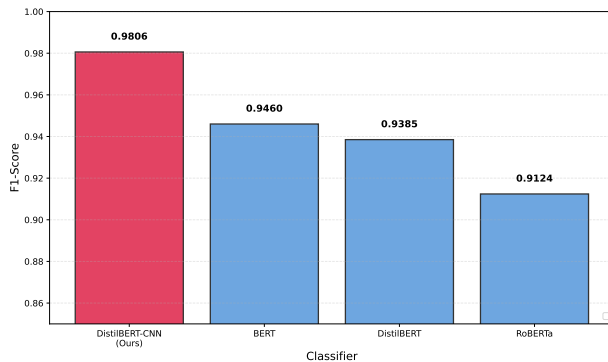


Table 6. Classification performance of the DistilBERT-CNN hybrid model on the original and LLM-augmented PHEME and MediaEval datasets.

Dataset	Augmentation Model	Accuracy	Precision	Recall	F1-Score
PHEME	None	0.8771	0.8765	0.8771	0.8767
	Gemma-2-2B	0.9704	0.9706	0.9704	0.9703
	LLaMA-3.2-3B	0.9720	0.9724	0.9720	0.9721
	DeepSeek-R1-Distill-Qwen-1.5B	0.9774	0.9774	0.9774	0.9774
	GPT-4o-mini	0.9805	0.9807	0.9805	0.9806
MediaEval	None	0.8016	0.8332	0.8016	0.8074
	Gemma-2-2B	0.8709	0.8864	0.8709	0.8738
	LLaMA-3.2-3B	0.8678	0.8779	0.8678	0.8702
	DeepSeek-R1-Distill-Qwen-1.5B	0.8871	0.8893	0.8871	0.8879
	GPT-4o-mini	0.9000	0.9048	0.9000	0.9013

Table 7. Statistical significance test results comparing GPT-4o-mini against other augmentation models.

Dataset	Model	Difference	p-value	Cohen's d
PHEME	None	0.1038	0.000005	14.95
	Gemma-2-2B	0.0133	0.005324	2.46
	LLaMA-3.2-3B	0.0120	0.005010	2.50
	DeepSeek-R1-Distill-Qwen-1.5B	0.0070	0.021898	1.63
MediaEval	None	0.0938	0.000015	11.26
	Gemma-2-2B	0.0310	0.000427	4.80
	LLaMA-3.2-3B	0.0353	0.000164	6.13
	DeepSeek-R1-Distill-Qwen-1.5B	0.0180	0.001480	3.47

**Figure 3.** Comparison of classifier performance on the PHEME dataset augmented with GPT-4o-mini.

4.3.4 Classifier Architecture Analysis

To further validate our architectural design choice, we compared the performance of our DistilBERT-CNN hybrid model against standard transformer-based classifiers on the PHEME dataset augmented with GPT-4o-mini. Figure 3 presents the comparative results.

We attribute this performance gain to the architectural synergy of our design: DistilBERT captures global contextual representations, while the CNN layers extract local n-gram patterns and position-invariant features that are particularly relevant for rumor detection. This hybrid approach leverages the strengths of both architectures, enabling the model to capture both long-range dependencies and local lexical cues. These findings confirm that our hybrid architecture yields substantial gains over standard transformer-based classifiers.

4.3.5 Comparison with Conventional Augmentation Methods

To comprehensively evaluate the effectiveness of our proposed LLM-based augmentation framework, we compared it against three representative baseline techniques for handling class imbalance:

- **Random Oversampling (ROS):** A standard imbalance mitigation technique that randomly



duplicates minority-class instances to balance the class distribution. While simple and computationally efficient, ROS can lead to overfitting as it does not introduce any new information, merely replicating existing samples.

- **Easy Data Augmentation (EDA):** A lightweight text augmentation method that applies a combination of simple lexical transformations, including synonym replacement, random insertion, random swap, and random deletion. EDA introduces moderate diversity while preserving the original meaning, making it a popular baseline for text classification tasks.
- **Back Translation (BT):** A paraphrasing technique where text is translated from the source language to an intermediate language and then translated back to the source language. This process generates semantically equivalent but lexically diverse paraphrases. We employed MarianMT pre-trained models and used French, German, and Spanish as intermediate languages to maximize linguistic diversity. Back translation preserves semantic meaning while introducing natural paraphrasing, making it a strong augmentation baseline.

All augmentation methods were applied exclusively to the training folds under the same 5-fold cross-validation protocol to ensure a fair comparison. The DistilBERT-CNN classifier was trained and evaluated under identical conditions for each baseline method.

Figure 4 presents the comparative classification results in terms of F1-score. Random Oversampling provides moderate improvements over the unaugmented baseline, confirming the importance of addressing class imbalance. EDA achieves slightly better performance than ROS, demonstrating that simple lexical transformations can enhance model generalization. Back Translation outperforms both ROS and EDA, indicating that semantically preserved paraphrasing offers more valuable training signal than simple duplication or lexical substitutions.

However, all three conventional methods are substantially outperformed by our LLM-based augmentation approach. GPT-4o-mini augmentation achieves F1-scores of 0.9806 on PHEME and 0.9013 on MediaEval, representing improvements of 10.3% and 9.4% over the baselines, respectively. This significant performance gap can be attributed to the ability of fine-tuned LLMs to generate semantically rich, contextually coherent, and stylistically authentic synthetic samples that closely mirror real-world rumor patterns, a capability that simpler augmentation techniques cannot replicate. Notably, even the smallest LLM-based augmentation (Gemma-2-2B) achieves comparable or

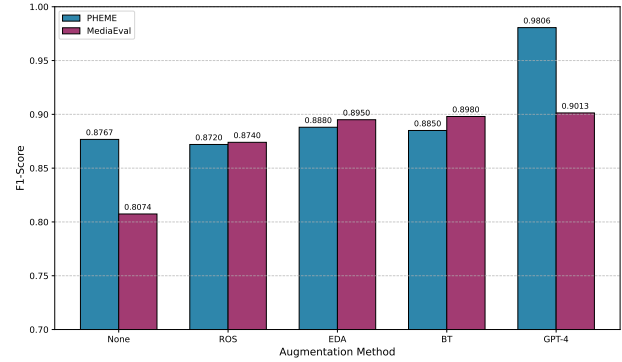


Figure 4. Comparison of F1-scores achieved by different augmentation methods on the PHEME and MediaEval datasets.

superior performance to the best conventional method, underscoring the value of generative approaches for data augmentation in rumor detection.

4.4 Evaluation of Generated Text

To assess the quality of the generated tweets beyond downstream classification performance, we conducted a comprehensive evaluation from multiple complementary perspectives. Rather than relying solely on classification accuracy, we evaluated whether the synthetic samples are fluent, semantically consistent with the original data distribution, sufficiently diverse, and capable of preserving the emotional and thematic characteristics of the target class. Our evaluation framework comprises five complementary analyses, each targeting a distinct dimension of synthetic data quality.

First, we measure linguistic fluency by computing GPT-2 perplexity, which quantifies how naturally structured and coherent the generated text is relative to real-world language patterns. Second, we assess semantic alignment with the original data distribution using MAUVE, which measures the distributional similarity between generated and real texts in the embedding space of a pretrained language model. Third, we evaluate lexical diversity through Self-BLEU to ensure that the augmentation model produces a rich variety of expressions rather than repeating a small set of generic templates. Fourth, we examine sentiment consistency using VADER to verify that the distribution of positive, neutral, and negative tweets in the augmented data matches that of real-world rumors, ensuring that augmentation does not introduce emotional bias. Fifth, we analyze thematic coherence via Latent Dirichlet Allocation (LDA) to confirm that the underlying thematic structure remains consistent after augmentation, preventing topical bias.

Together, these evaluations ensure that our synthetic tweets are grammatically fluent, semantically faithful, sufficiently diverse, emotionally consistent, and topically representative, thereby enhancing down-



stream rumor classification without distorting the original dataset’s characteristics. The following subsections detail each evaluation method and its corresponding results.

4.4.1 Perplexity as a Measure of Fluency

Perplexity is a widely used metric in natural language processing (NLP) that quantifies how well a language model predicts a given text sequence. A lower perplexity score indicates that the generated text is more fluent and coherent, as it aligns closely with natural language patterns. Perplexity is defined as:

$$PPL(W) = \exp \left(-\frac{1}{N} \sum_{i=1}^N \log P(w_i | w_{1:i-1}) \right) \quad (7)$$

where:

- $W = (w_1, w_2, w_3, \dots, w_n)$ represents a sequence of words,
- N is the total number of words in the sequence,
- $P(w_i | w_{1:i-1})$ is the probability assigned by the model to word w_i given the previous words.

A lower average perplexity across the dataset signifies that the augmented text is more naturally structured and comprehensible. In our evaluation, we compute perplexity by tokenizing each generated text sample, feeding it into GPT-2, and computing the model’s loss while predicting the text. The exponential of the loss function yields the perplexity score, which provides an indicator of linguistic fluency. This approach allows for quantitative comparison of fluency across datasets generated by different LLMs. Table 8 summarizes the average perplexity scores for each LLM on both the PHEME and MediaEval datasets.

Table 8. Average perplexity scores of the original and LLM-augmented PHEME and MediaEval datasets.

Augmentation Model	PHEME	MediaEval
None	100.68	134.47
DeepSeek-R1-Distill-Qwen-1.5B	23.44	124.22
Gemma-2-2B	60.66	122.88
LLaMA-3.2-3B	19.16	125.34
GPT-4o-mini	52.76	285.09

Table 8 reports average perplexity for the original (no augmentation) texts and for synthetic tweets generated by various LLMs on both the PHEME and MediaEval datasets. On PHEME, all augmentation models dramatically reduce perplexity compared to the un-augmented baseline (100.68). In particu-

lar, LLaMA-3.2-3B (19.16) and DeepSeek-R1-Distill-Qwen-1.5B (23.44) achieve the lowest perplexities, indicating they generate text highly consistent with the patterns learned from PHEME. Gemma-2-2B (60.66) and GPT-4o-mini (52.76) also improve fluency relative to the original tweets, though to a lesser extent.

On MediaEval, where tweets tend to be shorter and laden with URLs, only DeepSeek-R1-Distill-Qwen-1.5B (124.22) and Gemma-2-2B (122.88) offer modest perplexity improvements over the baseline (134.47), suggesting their outputs align more closely with the terse, URL-rich style of this dataset. LLaMA-3.2-3B (125.34) exhibits a similar fit, whereas GPT-4o-mini’s high perplexity (285.09) signals a substantial mismatch: its generative patterns favor verbose, context-rich prose and factual completeness, often introducing filler tokens or expanded phrasing that conflict with MediaEval’s clipped syntax and frequent hyperlinks. Consequently, it assigns very low probabilities to URL-heavy sequences, resulting in elevated surprisal. Nonetheless, GPT-4o-mini still achieves the highest classification performance ($F1 = 0.9013$), indicating that, even with stylistic divergence, its rich representations contribute to superior downstream detection. This contrast underscores the importance of balancing perplexity and classification efficacy when selecting augmentation models.

4.4.2 Semantic Fidelity and Diversity: MAUVE and Self-BLEU

While perplexity measures fluency and LDA captures topical coherence, they do not directly assess whether the generated texts are semantically similar to real human-written rumors, nor whether the generator suffers from mode collapse (i.e., producing near-identical outputs). To fill this gap, we adopt two complementary metrics:

MAUVE [38] measures the distributional similarity between the generated and reference texts in the embedding space of a pretrained language model. It computes a divergence frontier between the two sets of embeddings, yielding a score in $[0, 1]$, where higher values indicate that the synthetic samples closely match the semantic manifold of the real rumors. Unlike n-gram overlap metrics, MAUVE is sensitive to long-range semantic coherence, making it suitable for detecting subtle shifts in meaning.

Self-BLEU evaluates the diversity within a generated corpus. It is calculated as the average 4-gram BLEU score of each generated sentence against all other generated sentences in the same set. A lower Self BLEU score indicates higher lexical and syntactic diversity, confirming that the augmentation model is producing a wide variety of expressions rather than



repeating a small set of generic templates. Preserving diversity is crucial for rumor classification, as real misinformation spans a wide range of linguistic styles.

Table 9 reports the MAUVE and Self BLEU scores for each augmentation model on the PHEME and MediaEval datasets.

The results reveal noticeable differences among the four LLMs. On the PHEME dataset, GPT-4o-mini achieved the lowest Self-BLEU score (0.438), indicating that it generated the most diverse set of rumor tweets. It also obtained the highest MAUVE score (0.611), suggesting that the generated samples most closely matched the distribution of authentic rumor tweets. Gemma-2-2B ranked second in terms of MAUVE (0.591), while DeepSeek-R1-Distill-Qwen-1.5B and LLaMA-3.2-3B produced slightly less realistic distributions.

On the MediaEval dataset, however, the ranking differs substantially. DeepSeek-R1-Distill-Qwen-1.5B and LLaMA-3.2-3B achieved the highest MAUVE scores (0.953 and 0.952, respectively), demonstrating an excellent ability to reproduce the distribution of genuine tweets. Both models also obtained considerably lower Self-BLEU values (0.613) than Gemma-2-2B and GPT-4o-mini, indicating a favorable balance between diversity and realism. Although GPT-4o-mini achieved the best semantic quality on PHEME, it produced substantially higher Self-BLEU (0.914) and lower MAUVE (0.828) on MediaEval, suggesting that its generated real-news tweets were more repetitive and deviated further from the original data distribution.

4.4.3 Sentiment Analysis Method

To assess the emotional tone of the textual data, sentiment analysis is conducted using a lexicon-based approach. Specifically, we employ VADER (Valence Aware Dictionary and sEntiment Reasoner) [39], a rule-based model designed to capture sentiment polarity in social media text. VADER combines a sentiment lexicon with heuristics for syntactic and grammatical cues to generate sentiment scores.

The sentiment classification relies on the compound score derived from lexical features and intensity modifiers. This score, which ranges from -1 to +1, represents the overall sentiment of a given text. Based on empirically validated thresholds, text is categorized into three sentiment classes: positive, neutral, and negative. The distribution of these classes across the dataset provides insights into the prevailing emotional trends within the content.

Table 10 reports the counts of positive, neutral, and negative tweets for the original and LLM-augmented

PHEME and MediaEval datasets. In PHEME, the original distribution (1,521 positive; 1,948 neutral; 2,956 negative) exhibits a clear hierarchy of sentiment classes. Following augmentation, all models proportionally increase each sentiment category, DeepSeek-R1-Distill-Qwen-1.5B (1,845-2,270-3,931), Gemma-2-2B (1,863-2,230-3,953), LLaMA-3.2-3B (1,761-2,267-4,018), and GPT-4o-mini (1,680-2,348-4,018), yet the relative ordering (*negative* > *neutral* > *positive*) remains unchanged. This consistency indicates that the synthetic tweets preserve the original dataset's overall emotional balance.

For MediaEval, the original texts are overwhelmingly classified as neutral (11,851 neutral; 0 positive or negative), reflecting the short, URL-heavy content that VADER treats as sentiment-neutral. After augmentation, all models introduce both positive and negative tweets while maintaining a dominant neutral class: DeepSeek-R1-Distill-Qwen-1.5B (1,044-12,420-250), Gemma-2-2B (306-13,329-79), LLaMA-3.2-3B (401-13,038-241), and GPT-4o-mini (826-12,573-283). In each case, neutral tweets still comprise the vast majority, mirroring the original dataset's neutral bias.

Together, these results demonstrate that LLM-based augmentation expands the dataset without distorting its underlying sentiment profile. By preserving the original class proportions, negative-dominant in PHEME and neutral-dominant in MediaEval, the synthetic data maintains fidelity to real-world sentiment patterns, ensuring that downstream models learn from a representative emotional distribution.

4.4.4 Topic Modeling Evaluation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that uncovers latent thematic structure within a corpus. It assumes each document d is generated from a mixture of K topics, where each topic k is characterized by a distribution ϕ_k over the vocabulary, drawn from a Dirichlet prior $\text{Dirichlet}(\beta)$. Simultaneously, each document has its own topic mixture $\theta_d \sim \text{Dirichlet}(\alpha)$. For each word $w_{d,n}$ in document d , a topic assignment $z_{d,n}$ is sampled from $\text{Categorical}(\theta_d)$, and the word itself is sampled from $\text{Categorical}(\phi_{z_{d,n}})$. Inference, typically via variational methods or Gibbs sampling, yields estimates of θ_d (document–topic proportions) and ϕ_k (topic–word distributions), allowing identification of the most salient words per topic.

To ensure a principled selection of the number of topics, we performed a systematic grid search over $K \in \{5, 8, 10, 12, 15, 20\}$ using topic coherence (Cv) as the objective metric. Based on this analysis, we selected $K = 10$ for both datasets, as this configuration yielded the highest coherence scores while maintaining interpretable topic distinctions. For the original



Table 9. MAUVE (semantic similarity) and Self BLEU (diversity) for LLM-augmented PHEME and MediaEval datasets.

Dataset	LLM	Self-BLEU ↓	MAUVE ↑
PHEME	DeepSeek-R1-Distill-Qwen-1.5B	0.453	0.578
	Gemma-2-2B	0.445	0.591
	LLaMA-3.2-3B	0.441	0.569
	GPT-4o-mini	0.438	0.611
MediaEval	DeepSeek-R1-Distill-Qwen-1.5B	0.613	0.953
	Gemma-2-2B	0.662	0.931
	LLaMA-3.2-3B	0.613	0.952
	GPT-4o-mini	0.914	0.828

Table 10. Comparison of sentiment distribution on the original and LLM-augmented PHEME and MediaEval datasets.

Dataset	Model	Positive	Neutral	Negative
PHEME	None	1521	1948	2956
	DeepSeek-R1-Distill-Qwen-1.5B	1845	2270	3931
	Gemma-2-2B	1863	2230	3953
	LLaMA-3.2-3B	1761	2267	4018
	GPT-4o-mini	1680	2348	4018
MediaEval	None	0	11851	0
	DeepSeek-R1-Distill-Qwen-1.5B	1044	12420	250
	Gemma-2-2B	306	13329	79
	LLaMA-3.2-3B	401	13038	241
	GPT-4o-mini	826	12573	283

PHEME dataset, the average topic coherence (Cv) was 0.52, while for MediaEval it was 0.48. After augmentation, coherence scores remained stable across all LLM-generated datasets, ranging from 0.49 to 0.53, confirming that the latent thematic structure was preserved and not artificially distorted by the augmentation process.

Tables 11 (PHEME) and 12 (MediaEval) compare the top LDA-extracted topics before and after augmentation with various LLMs. In both corpora, the original datasets feature event- or context-specific keywords, such as “ferguson,” “charliehebdo,” and “sydneyseige” in PHEME, and “sandy,” “hurricane,” and “rt” in MediaEval. After augmentation with DeepSeek-R1-Distill-Qwen-1.5B, Gemma-2-2B, LLaMA-3.2-3B, and GPT-4o-mini, these core terms remain dominant. Although each model produces slight shifts in word prominence (e.g., DeepSeek-R1-Distill-Qwen-1.5B marginally increases the weight of “attack” and

“hebdo,” while LLaMA-3.2-3B elevates “paris” and “freeexpression”), the fundamental topic structure is preserved. This indicates that LLM-generated synthetic tweets faithfully replicate the latent themes of the original data.

4.5 Ablation Study

To dissect the contribution of our design choices while respecting computational constraints, we conduct a targeted ablation study using our best-performing model (GPT-4o-mini) on the PHEME dataset. We evaluate three critical factors: (i) the inclusion of event-specific hashtags in the prompt, (ii) low temperature ($T = 0.1$), and (iii) high temperature ($T = 1.0$). The full model uses $T = 0.4$ (as per Table 5) and $\text{top-}p = 0.95$. We report downstream F1-score alongside MAUVE (semantic alignment) and Self-BLEU (diversity) to capture the full impact in Table 13.

Impact of Hashtags. Removing hashtag condi-



Table 11. Topic Distribution on the Original and LLM-augmented PHEME.

Augmentation Model	Top Topics
None	0.034*”ferguson” + 0.017*”police” + 0.019*”ottawa” + 0.018*”charliehebdo” + 0.015*”germanwings” + 0.014*”attack” + 0.011*”shooting”; 0.050*”sydneyneysiege” + 0.047*”sydney” + 0.029*”hostages” + 0.028*”cafe” + 0.024*”police”
DeepSeek-R1-Distill-Qwen-1.5B	0.046*”ferguson” + 0.034*”police” + 0.021*”officer” + 0.018*”brown” + 0.017*”michael” + 0.039*”sydneyneysiege” + 0.026*”sydney” + 0.021*”s” + 0.011*”place” + 0.010*”police”; 0.057*”hebdo” + 0.051*”charlie” + 0.036*”charliehebdo” + 0.031*”attack” + 0.028*”paris”
Gemma-2-2B	0.038*”police” + 0.032*”ferguson” + 0.031*”charlie” + 0.028*”hebdo” + 0.022*”charliehebdo”; 0.021*”s” + 0.011*”breaking” + 0.011*”ottawa” + 0.007*”russia” + 0.007*”charliehebdo”; 0.029*”sydney” + 0.025*”sydneyneysiege” + 0.014*”siege” + 0.009*”germanwings”
GPT-4o-mini	0.026*”attack” + 0.020*”update” + 0.019*”charliehebdo” + 0.017*”hebdo” + 0.017*”police”; 0.061*”police” + 0.042*”shooting” + 0.038*”ferguson” + 0.021*”man” + 0.016*”officer”; 0.029*”ottawa” + 0.015*”beng” + 0.014*”germanwings” + 0.011*”charliehebdo”
LLaMA-3.2-3B	0.055*”charlie” + 0.051*”hebdo” + 0.040*”paris” + 0.032*”charliehebdo” + 0.031*”attack”; 0.036*”sydney” + 0.022*”freexpression” + 0.021*”gunman” + 0.021*”sydneyneysiege” + 0.015*”police”; 0.044*”ferguson” + 0.033*”police” + 0.018*”shooting” + 0.016*”” + 0.016*”officer”

Table 12. Topic Distribution on the Original and LLM-augmented MediaEval.

Augmentation Model	Top Topics
None	0.101*”sandy” + 0.034*”hurricane” + 0.021*”rt” + 0.021*”hurricanesandy” + 0.021*”nyc”; 0.076*”sandy” + 0.032*”de” + 0.015*”la” + 0.012*”cuba” + 0.012*”rt” + 0.019*”boston” + 0.017*”sochi” + 0.017*”suspect” + 0.015*”mh370”
DeepSeek-R1-Distill-Qwen-1.5B	0.000*”jessica_lee” + 0.000*”mattbuckley13” + 0.000*”jessicaschwarz” + 0.000*”holy”; 0.075*”m” + 0.053*”new” + 0.050*”york” + 0.047*”happy” + 0.046*”city”; 0.050*”photo” + 0.032*”via” + 0.026*”rt” + 0.019*”boston” + 0.018*”“
Gemma-2-2B	0.079*”cover” + 0.066*”rt” + 0.065*”new” + 0.057*”york” + 0.051*”amazing”; 0.056*”sandy” + 0.049*”rt” + 0.032*”jersey” + 0.021*”power” + 0.019*”hurricane” + 0.000*”wrong” + 0.000*”result” + 0.000*”awesome” + 0.000*”cool”
GPT-4o-mini	0.051*”rt” + 0.042*”sandy” + 0.029*”boston” + 0.025*”suspect” + 0.021*”hurricane” + 0.000*”crazy” + 0.000*”sad” + 0.000*”8pm” + 0.000*”prayforamerica” 0.094*”rt” + 0.043*”new” + 0.041*”york” + 0.041*”cover” + 0.039*”magazine”
LLaMA-3.2-3B	0.064*”via” + 0.053*”sandy” + 0.040*”nymag” + 0.033*”rt” + 0.012*”staten” + 0.012*”island” + 0.011*”ave” + 0.010*”suspect” + 0.009*”mob” + 0.000*”jeffreygoldberg” + 0.000*”fake” + 0.000*”favorites”



Table 13. Targeted Ablation Study on GPT-4o-mini Generations over the PHEME Dataset.

Configuration	F1-score $\uparrow$	MAUVE $\uparrow$	Self-BLEU $\downarrow$
Full Model (Ours, $T = 0.4$)	0.9806	0.611	0.438
w/o Hashtag Conditioning	0.9428	0.461	0.538
Low Temperature ($T = 0.1$)	0.9756	0.571	0.558
High Temperature ($T = 1.0$)	0.9656	0.511	0.358

tioning causes the largest degradation, dropping the F1-score by 3.8 points (from 0.9806 to 0.9428). The MAUVE score plummets from 0.611 to 0.461, indicating that without hashtags, the generated tweets drift substantially away from the semantic manifold of real rumor data. Concurrently, Self-BLEU rises from 0.438 to 0.538, revealing that the model falls back on more repetitive templates. This confirms that hashtags are not superficial metadata but crucial semantic anchors; they explicitly ground the generator in specific event contexts (e.g., “Ferguson” vs. “Charlie Hebdo”), which is critical for producing class-consistent synthetic samples that preserve classification performance.

Impact of Temperature. Varying the temperature exposes a clear fluency-diversity trade-off. At low temperature ($T = 0.1$), the model produces conservative, highly predictable outputs. While this achieves competitive F1 (0.9756) and preserves reasonable semantic alignment (MAUVE 0.571), it results in the highest Self-BLEU (0.558), indicating undesirable lexical repetition. This suggests that overly deterministic sampling limits the variety of training examples, potentially harming classifier robustness despite maintaining high fluency.

Conversely, high temperature ($T = 1.0$) maximizes diversity (Self-BLEU drops to 0.358) but introduces excessive semantic noise, degrading MAUVE to 0.511 and F1 to 0.9656. The loss in semantic coherence directly explains the downstream performance drop. Our chosen intermediate temperature ($T = 0.4$) strikes the optimal balance. It preserves the semantic fidelity of the original data (MAUVE 0.611) while maintaining sufficient diversity (Self-BLEU 0.438) to avoid mode collapse, yielding the highest classification performance (0.9806).

Overall, this targeted ablation demonstrates that the effectiveness of our framework is not attributable to a single trivial choice, but rather to the synergistic combination of hashtag-guided context and carefully calibrated sampling stochasticity. These results robustly justify our architectural decisions and confirm

that each design component contributes meaningfully to the final performance.

4.6 Computational Cost and Scalability Analysis

To assess the practical applicability of our proposed augmentation framework, we report the computational costs, training times, inference times, and memory requirements associated with fine-tuning and generating synthetic data using each of the four LLMs. Understanding these costs is essential for evaluating the feasibility of deploying such systems in real-world or resource-constrained environments.

4.6.1 Experimental Setup

All fine-tuning and generation experiments for open-weight models were conducted on a single NVIDIA Tesla P100 GPU with 16 GB of VRAM, available through the Kaggle platform. The LoRA-based fine-tuning with 4-bit quantization was implemented using the Unsloth framework, which significantly reduces memory consumption and training time compared to full fine-tuning. GPT-4o-mini fine-tuning and generation were performed using the OpenAI supervised fine-tuning API, with costs reported in United States Dollars (USD) based on OpenAI’s pricing as of February 2026.

4.6.2 Fine-Tuning Time and Cost

Table 14 summarizes the fine-tuning time, inference time, and estimated costs for all four models. The open-weight models were fine-tuned for 50 optimization steps, consistent with the configuration described in Section 4.3. Training durations ranged from 1.8 hours (Gemma-2-2B) to 2.5 hours (DeepSeek-R1-Distill-Qwen-1.5B) on the Tesla P100 GPU. With Kaggle’s free GPU usage, the monetary cost for fine-tuning open-weight models is negligible.

For GPT-4o-mini, fine-tuning was conducted through the OpenAI platform at a cost of \$45.00 per 1M training tokens. The total fine-tuning cost for the PHEME dataset (1,746 training samples) was approximately \$78.57, while the MediaEval dataset (5,008 training samples) incurred approximately \$225.36. These costs reflect the supervised fine-tuning service provided by OpenAI, which includes infrastructure management and scalability.

4.6.3 Inference Time

Synthetic data generation time is another critical factor for practical deployment. For open-weight models, generation of all synthetic tweets (1,621 for PHEME,



1,832 for MediaEval) was completed in approximately 1.5–2.0 seconds per tweet on the Tesla P100 GPU. GPT-4o-mini generation was faster, averaging 0.8 seconds per tweet, though this incurs API costs. Total generation time for each dataset was:

- **DeepSeek-R1-Distill-Qwen-1.5B:** 49 minutes (PHEME), 55 minutes (MediaEval)
- **Gemma-2-2B:** 41 minutes (PHEME), 46 minutes (MediaEval)
- **LLaMA-3.2-3B:** 43 minutes (PHEME), 49 minutes (MediaEval)
- **GPT-4o-mini:** 22 minutes (PHEME), 24 minutes (MediaEval) (API-based)

4.6.4 Memory Requirements

Memory consumption during fine-tuning is a key consideration for researchers with limited hardware resources. With 4-bit quantization and LoRA fine-tuning, all open-weight models fit comfortably within the 16 GB VRAM of the Tesla P100:

- **DeepSeek-R1-Distill-Qwen-1.5B (671B MoE):** 18 GB VRAM (requires gradient checkpointing)
- **Gemma-2-2B (2B):** 14 GB VRAM
- **LLaMA-3.2-3B (3B):** 16 GB VRAM

4.6.5 Scalability Considerations

Several factors influence the scalability of our proposed approach to larger datasets:

Open-Weight Models (DeepSeek-R1-Distill-Qwen-1.5B, Gemma-2-2B, LLaMA-3.2-3B):

- **Training Scalability:** LoRA-based fine-tuning scales linearly with dataset size. For datasets up to 50,000 samples, training remains feasible on a single Tesla P100 GPU. For larger datasets, distributed training or model parallelism would be required.
- **Generation Scalability:** Inference can be efficiently batched to reduce per-sample time. With batch sizes of 8–16, generation throughput can be increased by 2 – 3× on the Tesla P100.
- **Memory Scalability:** The 4-bit quantization used in this study reduces memory requirements by approximately 75% compared to full-precision models. For larger MoE models, offloading to CPU memory or using multiple GPUs would be necessary.

GPT-4o-mini (OpenAI API):

- **Training Scalability:** OpenAI's platform automatically scales to handle large datasets, with no infrastructure management required. Cost scales

linearly with the number of training tokens.

- **Generation Scalability:** API-based generation can be parallelized across multiple requests, though rate limits may apply depending on the service tier.
- **Cost Scalability:** While convenient, the pay-per-token model becomes expensive for large-scale applications (e.g., augmenting datasets with millions of samples). This cost must be weighed against the performance benefits.

4.7 Discussion

The experimental evaluation presented in this section yields several insights that extend beyond the immediate performance comparisons and merit deeper discussion. We organize our analysis around three interconnected themes: the fundamental advantage of generative augmentation over conventional methods, the dataset-dependent nature of LLM effectiveness, and the implications of our quality assessment findings for synthetic data evaluation.

Perhaps the most significant finding of this study is the substantial gap between conventional augmentation techniques and LLM-based generation. This performance disparity is not merely quantitative but qualitative in nature. Random Oversampling, Easy Data Augmentation, and Back Translation operate at the surface level of text, duplicating, substituting synonyms, or paraphrasing through translation. These methods, while computationally efficient, are fundamentally constrained by the information already present in the training data. They cannot introduce novel semantic patterns, stylistic variations, or contextual nuances that are not already represented in the original corpus. In contrast, fine-tuned LLMs learn the underlying generative distribution of the target class, enabling them to produce samples that, while synthetic, capture the complex interplay of lexical choice, syntactic structure, and contextual framing that characterizes authentic rumor discourse. This capacity to generate genuinely novel training examples, rather than merely recombining existing ones, explains why LLM-based augmentation consistently outperforms conventional methods by such a wide margin. The magnitude of this gap, particularly on PHEME where GPT-4o-mini achieves a relative improvement of over 10% compared to the best conventional method, underscores the fundamental difference between surface-level text manipulation and deep generative modeling of linguistic patterns.

The dataset-specific variation in LLM performance reveals a more nuanced picture than a simple ranking of model capabilities. On PHEME, GPT-4o-mini's dominance is unambiguous across all metrics: high-



Table 14. Computational Cost and Runtime Comparison of LLMs Used for Data Augmentation.

Model	Hardware	Training Time	Inference Time	Memory	Estimated Cost
DeepSeek-R1-Distill-Qwen-1.5B	Kaggle Tesla P100	~2.5 hours	~1.8 sec/tweet	~18 GB	~\$0
Gemma-2-2B	Kaggle Tesla P100	~1.8 hours	~1.5 sec/tweet	~14 GB	~\$0
GPT-4o-mini	OpenAI API	~3–4 hours	~0.8 sec/tweet	N/A	\$45.00 / 1M tokens
LLaMA-3.2-3B	Kaggle Tesla P100	~2.0 hours	~1.6 sec/tweet	~16 GB	~\$0

est F1-score, highest MAUVE, and lowest Self-BLEU. This suggests that PHEME rumors, which are event-driven, context-rich narratives, benefit from GPT-4o-mini’s capacity for coherent, elaborated generation. The model’s ability to maintain narrative consistency and produce stylistically varied outputs aligns well with the linguistic complexity of authentic rumors. However, on MediaEval, the landscape becomes more complex. Here, open-weight models achieve superior semantic alignment (MAUVE) and diversity (Self-BLEU), yet GPT-4o-mini still yields the best classification performance. This apparent contradiction highlights a critical distinction: semantic alignment with the original distribution does not always translate to optimal training signal for classification. MediaEval tweets are short, URL-heavy, and information-sparse. GPT-4o-mini’s elaborated, context-rich outputs, while stylistically divergent, provide the classifier with richer semantic features to learn from. The model trades distributional fidelity for informational richness, and in this specific context, that trade-off benefits classification performance. This finding suggests that the optimal augmentation strategy is not universal but depends on the characteristics of the target dataset, particularly its linguistic richness and stylistic homogeneity.

Our multi-dimensional quality assessment reveals important limitations of relying on any single metric to evaluate synthetic data. Perplexity, while a standard measure of fluency, proved to be a poor predictor of classification performance in certain contexts. GPT-4o-mini’s high perplexity on MediaEval did not indicate low-quality data; rather, it reflected a beneficial departure from the terse, URL-heavy style of the original dataset. MAUVE and Self-BLEU provided complementary information, revealing that GPT-4o-mini’s MediaEval outputs, while semantically rich, exhibited lower diversity than those of other models. This suggests that the ideal synthetic dataset may require a delicate balance: sufficient fluency to be readable, sufficient semantic alignment to be realistic, sufficient diversity to avoid mode collapse, and sufficient informational richness to provide useful training signals. No single metric captures all these dimensions, and the relative importance of each may vary across datasets and tasks. This has important implica-

tions for the design of augmentation pipelines: quality assessment should be multi-faceted and task-aware, with metrics selected based on the specific characteristics of the target dataset and the downstream task requirements. The complementary nature of our evaluation framework, combining perplexity, MAUVE, Self-BLEU, VADER, and LDA, demonstrates that a holistic approach to quality assessment is essential for understanding the true value of synthetic data for downstream tasks.

The practical implications of these findings are twofold. First, practitioners should not assume that a model’s performance on one dataset will generalize to others; model selection should be informed by the linguistic characteristics of the target data. Second, quality assessment should go beyond fluency metrics to include semantic alignment, diversity, and task-relevant characteristics. These insights collectively advance our understanding of how to effectively leverage LLMs for data augmentation in misinformation detection and provide guidance for future research in this domain.

5 Conclusions

In this work, we addressed the persistent challenge of class imbalance in rumor and fake-news detection by introducing a multi-model LLM-based data augmentation framework. Our approach leverages four state-of-the-art large language models fine-tuned on the minority classes of two benchmark datasets, PHEME and MediaEval. Specifically, the models were adapted to generate rumor tweets for PHEME and real tweets for MediaEval through class-conditional instruction-based fine-tuning. This strategy enabled each model to capture the linguistic, stylistic, and contextual characteristics of the target minority class, producing synthetic samples that enrich underrepresented classes while preserving the essential properties of authentic social media discourse.

To ensure the reliability of the generated data, we employed a comprehensive multi-dimensional quality validation pipeline. This included GPT-2 perplexity for linguistic fluency, VADER for sentiment alignment, LDA for thematic coherence, MAUVE for semantic similarity to the original data distribution, and



Self-BLEU for diversity assessment. Together, these complementary metrics confirmed that the synthetic tweets remained fluent, semantically faithful, diverse, emotionally consistent, and topically representative of the original corpora.

When incorporated into a DistilBERT-CNN hybrid classifier, the augmented datasets produced substantial and consistent performance gains across both benchmarks. On PHEME, the F1-score improved from 0.8767 to 0.9806, while accuracy increased from 0.8771 to 0.9805, precision from 0.8765 to 0.9807, and recall from 0.8771 to 0.9805. On MediaEval, the F1-score rose from 0.8074 to 0.9013, with accuracy improving from 0.8016 to 0.9000, precision from 0.8332 to 0.9048, and recall from 0.8016 to 0.9000. These improvements were statistically validated through paired t-tests, with $p < 0.05$ and large effect sizes ($Cohen's d > 0.8$). Moreover, the proposed LLM-based augmentation consistently outperformed conventional class-imbalance techniques, including Random Oversampling, Easy Data Augmentation, and Back Translation, which achieved substantially lower F1-scores across both datasets.

An ablation study using the best-performing model, GPT-4o-mini, on the PHEME dataset confirmed the importance of our design choices. Hashtag conditioning proved critical for semantic fidelity: removing it caused the largest performance drop, reducing the F1-score from 0.9806 to 0.9428 and MAUVE from 0.611 to 0.461. Temperature analysis revealed a clear diversity-coherence trade-off. Low temperature ($T = 0.1$) produced repetitive outputs with high Self-BLEU, while high temperature ($T = 1.0$) increased diversity but degraded semantic alignment and downstream F1. The selected intermediate temperature ($T = 0.4$) achieved the best balance, preserving semantic coherence while maintaining sufficient diversity to avoid mode collapse.

From a practical standpoint, the computational cost analysis demonstrates that the proposed framework is accessible across diverse resource settings. The open-weight models were fine-tuned on a single NVIDIA Tesla P100 GPU with 14–18 GB of VRAM in approximately 1.8–2.5 hours, at negligible monetary cost. GPT-4o-mini achieved the best classification performance through the OpenAI fine-tuning API, though at additional cost. These results indicate that LLM-based augmentation can be effectively deployed both in resource-constrained academic settings and in high-stakes applications where API-based models are feasible.

Limitations

Several limitations of our study should be acknowledged. First, our experiments were conducted on English-language datasets from Twitter, limiting generalizability to other languages, platforms, or misinformation types. Second, we focused exclusively on text-based augmentation, leaving multimodal scenarios, where rumors propagate through images, videos, and text, as an important extension. Third, the computational cost analysis reflects the specific hardware and API pricing available at the time of writing; these factors are subject to change.

Future Directions

Our findings open multiple avenues for future research. First, we plan to extend our framework to multimodal settings by incorporating images and videos, leveraging recent advances in multimodal LLMs to generate rich, cross-modal synthetic data for misinformation detection. Second, we aim to develop adaptive generation techniques that can respond to evolving misinformation patterns in real time, potentially through online fine-tuning or reinforcement learning. Third, cross-lingual fine-tuning and augmentation present a valuable direction for supporting low-resource languages and global misinformation detection efforts. Fourth, we intend to develop task-aware quality metrics that better predict downstream classification performance, moving beyond standard fluency and diversity measures.

In conclusion, our work establishes that fine-tuned LLM-based augmentation is a robust, scalable, and accessible solution for addressing class imbalance in rumor and fake-news detection. The substantial performance gains over conventional methods, the validation of our design choices through ablation, and the practical feasibility of open-weight models collectively position our framework as a valuable tool for real-world misinformation detection systems. As misinformation continues to evolve in complexity and scale, we believe that LLM-powered augmentation will play an increasingly important role in building more robust, adaptable, and reliable detection systems.

References

- [1] N. Difonzo and P. Bordia. *Rumor Psychology: Social and Organizational Approaches*. American Psychological Association, 2007. doi:[10.1037/11503-000](https://doi.org/10.1037/11503-000).
- [2] G.W. Allport and Leo Postman. The psychology of rumor. *Psychological Review*, 1947.
- [3] M. Askarizade. Enhancing rumor detection with



- data augmentation and generative pre-trained transformer. *Expert Systems Applications*, 262: 125649, 2025. doi:[10.1016/J.ESWA.2024.125649](https://doi.org/10.1016/J.ESWA.2024.125649).
- [4] Mojgan Askarizade, Ensieh Davoodijam, and Behrouz Tork Ladani. Advancing rumor detection in social media: A quality-stratified systematic review from cnns to llm-gnn hybrids. *Array*, page 100842, 2026. doi:[10.1016/j.array.2026.100842](https://doi.org/10.1016/j.array.2026.100842).
 - [5] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*, pages 675–684, 2011. doi:[10.1145/1963405.1963500](https://doi.org/10.1145/1963405.1963500).
 - [6] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. Prominent features of rumor propagation in online social media. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 1103–1108, 2013. doi:[10.1109/ICDM.2013.61](https://doi.org/10.1109/ICDM.2013.61).
 - [7] Hardeep Kaur Thakur, Aditi Gupta, Anshika Bhardwaj, and Dinesh Verma. Rumor detection on twitter using a supervised machine learning framework. *International Journal of Information Retrieval Research*, 8(3):1–13, 2018. doi:[10.4018/IJIRR.2018070101](https://doi.org/10.4018/IJIRR.2018070101).
 - [8] Shihang Wang, Zhen Li, Yumeng Wang, and Qiang Zhang. Machine learning methods to predict social media disaster rumor refuters. *International Journal of Environmental Research and Public Health*, 16(8):1452, 2019. doi:[10.3390/ijerph16081452](https://doi.org/10.3390/ijerph16081452).
 - [9] Xiaoting He, Gulila Tuerhong, Mairidan Wushouer, and Dong Xin. Rumors detection based on lifelong machine learning. *IEEE Access*, 10:25605–25620, 2022. doi:[10.1109/ACCESS.2022.3152842](https://doi.org/10.1109/ACCESS.2022.3152842).
 - [10] J. Ma, W. Gao, P. Mitra, S. Kwon, B.J. Jansen, K.-F. Wong, and M. Cha. Rumor detection on twitter. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 3818–3824, 2016.
 - [11] N. Ruchansky, S. Seo, and Y. Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 797–806, 2017. doi:[10.1145/3132847.3132877](https://doi.org/10.1145/3132847.3132877).
 - [12] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 795–816, 2017.
 - [13] Y. Peng and J. Wang. Rumor detection based on attention cnn and time series of context information. *Futur. Internet*, 13, 2021. doi:[10.3390/fi13110267](https://doi.org/10.3390/fi13110267).
 - [14] Sachin Kumar, R. Asthana, S. Upadhyay, N. Upreti, and M. Akbar. Fake news detection using deep learning models: A novel approach. *Trans Emerging Telecommun Technol*, 2019. doi:<https://doi.org/10.1002/ett.3767>.
 - [15] N. Rani, P. Das, and A.K. Bhardwaj. A hybrid deep learning model based on cnn-bilstm for rumor detection. In *Proceedings of the 6th International Conference on Communications and Electronics Systems*, pages 1423–1427, 2021. doi:[10.1109/ICCSE51350.2021.9489214](https://doi.org/10.1109/ICCSE51350.2021.9489214).
 - [16] D. Choi, H. Oh, S. Chun, T. Kwon, and J. Han. Preventing rumor spread with deep learning. *Expert Systems Applications*, 197:116688, 2022. doi:[10.1016/J.ESWA.2022.116688](https://doi.org/10.1016/J.ESWA.2022.116688).
 - [17] J. Ma, J. Li, W. Gao, Y. Yang, and K.F. Wong. Improving rumor detection by promoting information campaigns with transformer-based generative adversarial learning. *IEEE Transactions on Knowledge and Data Engineering*, 35:2657–2670, 2023. doi:[10.1109/TKDE.2021.3112497](https://doi.org/10.1109/TKDE.2021.3112497).
 - [18] Q. Chang, X. Li, and Z. Duan. A novel approach for rumor detection in social platforms: Memory-augmented transformer with graph convolutional networks. *Knowledge-Based Systems*, 292:111625, 2024. doi:[10.1016/J.KNOSYS.2024.111625](https://doi.org/10.1016/J.KNOSYS.2024.111625).
 - [19] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: Synthetic minority over-sampling technique — journal of artificial intelligence research. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.
 - [20] R. Akter, M.M. Rahman, A. Debnath, N. Samrat, M. Roman, A.N. Khan, N. Alvi, and M. Raihan. Enhancing rumor detection of twitter tweets using a deep learning approach. In *Proceedings of the 2024 15th International Conference on Computer Communications and Network Technology*, 2024. doi:[10.1109/ICCCNT61001.2024.10724559](https://doi.org/10.1109/ICCCNT61001.2024.10724559).
 - [21] Z. Zojaji and B. Tork Ladani. Adaptive cost-sensitive stance classification model for rumor detection in social networks. *Social Network Analysis and Mining*, 12:1–17, 2022. doi:[10.1007/S13278-022-00952-2/TABLES/4](https://doi.org/10.1007/S13278-022-00952-2/TABLES/4).
 - [22] X. Chen, D. Zhu, D. Lin, and D. Cao. Rumor knowledge embedding based data augmentation for imbalanced rumor detection. *Information Sciences*, 580:352–370, 2021. doi:[10.1016/J.INS.2021.08.059](https://doi.org/10.1016/J.INS.2021.08.059).
 - [23] DeepSeek-AI et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*, pages 1–22, 2025. doi:[arXiv:2501.12948](https://arxiv.org/abs/2501.12948).
 - [24] M. Riviere, S. Pathak, P.G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, and etc. Gemma 2: Improving open language models at a practical size. 2024.



- doi:[arXiv:2408.00118](https://arxiv.org/abs/2408.00118).
- [25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Ilya Sutskever, and Dario Amodei. Language models are unsupervised multitask learners. *OpenAI Blog*, 1, 2019. doi:[2007.07582](https://arxiv.org/abs/2007.07582).
 - [26] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, and etc. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL <https://arxiv.org/abs/2303.08774>.
 - [27] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, and etc. The llama 3 herd of models. *ArXiv preprint*, 2024. doi:[arXiv:2407.21783](https://arxiv.org/abs/2407.21783).
 - [28] J. Lai, X. Yang, W. Luo, L. Zhou, L. Li, Y. Wang, and X. Shi. Rumorllm: A rumor large language model-based fake-news-detection data-augmentation approach. *Applied Sciences*, 14: 3532, 2024. doi:[10.3390/AP14083532](https://doi.org/10.3390/AP14083532).
 - [29] Razieh Chalehchaleh, Reza Farahbakhsh, and Noel Crespi. Addressing data scarcity in multilingual fake news detection: an llm-based dataset augmentation approach. *Social Network Analysis and Mining*, 15, 2025. doi:[10.1007/s13278-025-01505-z](https://doi.org/10.1007/s13278-025-01505-z).
 - [30] Zhao Tong, Yimeng Gu, Huidong Liu, Qiang Liu, Shu Wu, Haichao Shi, and Xiao-Yu Zhang. Generate first, then sample: Enhancing fake news detection with llm-augmented reinforced sampling. volume 1, page 24276–24290, 2025. doi:[10.18653/v1/2025.acl-long.1182](https://doi.org/10.18653/v1/2025.acl-long.1182).
 - [31] Herun Wan, Jiaying Wu, Minnan Luo, Zhi Zeng, and Zhixiong Su. Truth over tricks: Measuring and mitigating shortcut learning in misinformation detection. *Advances in Neural Information Processing Systems*, 38:141018–141060, 2026. doi:[10.52202/085713-4712](https://doi.org/10.52202/085713-4712).
 - [32] Xiaojie Zhou, Tong Ning, Wang Deguang, and Li Fengqi. From latent structures to explicit reasoning: Synergizing clustering and llm for explainable and robust rumor detection. *Information Processing and Management*, 63, 2026. doi:[10.1016/j.ipm.2026.104872](https://doi.org/10.1016/j.ipm.2026.104872).
 - [33] Jiran Tao, Wang Cheng, and Jiang Binyan. Llm-enhanced rumor detection via virtual node induced edge prediction. *arXiv preprint*, 2026. URL [arXiv:2602.13279](https://arxiv.org/abs/2602.13279).
 - [34] Ziyu Luo, Wei Li, Haoping Huang, Kecheng Liu, and Min Gao. Sharp: Soft and hard prompt-guided augmentation with llm for low resource fake news detection. *Expert Systems with Applications*, page 131178, 2026. doi:[10.1016/j.eswa.2026.131178](https://doi.org/10.1016/j.eswa.2026.131178).
 - [35] Victor Sanh, Lysandre Debut, Julien Chaumond, et al. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv preprint*, 2024. doi:[1910.01108](https://doi.org/10.1910.01108).
 - [36] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, and P. Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11, 2016.
 - [37] Christina Boididou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, and Yannis Kompatsiaris. Verifying multimedia use at mediaeval 2015. In *MediaEval 2015 Workshop*, 2015.
 - [38] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *NeurIPS*, 2021.
 - [39] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014. doi:[10.1609/icwsn.v8i1.14550](https://doi.org/10.1609/icwsn.v8i1.14550).



Dr. Mojgan Askarizade is an Assistant Professor in the Department of Computer Engineering at Ardakan University, Iran. She received her B.Sc. degree in Computer Engineering from Yazd University in 2008, followed by an M.Sc. degree in Computer Engineering, with a specialization in Software Engineering, from the University of Isfahan in 2012. She obtained her Ph.D. degree in Computer Engineering in 2021. Her research interests primarily focus on Natural Language Processing (NLP), Data Mining, Neural Networks, Deep Learning, and Large Language Models (LLMs). Her research particularly explores the application of machine learning and deep learning techniques to text analysis, information extraction, rumor and misinformation detection, and the development of intelligent models for processing and understanding textual data.

